SCHOOL OF PHARMACY

Computational studies of Glucocerebrosidase in complex with its facilitator protein Saposin-C.

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Raquel Romero

The research Department of Pharmaceutical and Biological Chemistry,

UCL School of Pharmacy.

First supervisor: **DR. SHOZEB HAIDER**, UCL SCHOOL OF PHARMACY.

Second supervisor: **PROF. STEPHEN NEIDLE**, UCL SCHOOL OF PHARMACY.

PLAGIARISM STATEMENT

This thesis describes research conducted at the UCL School of Pharmacy between February of 2014 and September of 2017 under the supervision of Dr. Shozeb Haider. I certify that the research described is original. I also certify that I have written all the text herein and have clearly indicated by suitable citation any part of this dissertation that has already been published.

Raquel Romero Date:

The research Department of Pharmaceutical and Biological Chemistry, UCL School of Pharmacy.

ABSTRACT

Gaucher's Disease (GD) is a rare recessive disorder produced by the dysfunction of the lysosomal enzyme Glucocerebrosidase (GCase). GCase catalyses the cleavage of the glycolipid Glucosylceramide. The lack of functional GCase leads to the accumulation of its lipid substrate in lysosomes causing GD. GD presents a great phenotypic variation, symptoms ranging from asymptomatic adults to early childhood death due to neurological damage. More than 250 mutations in the protein GCase have been discovered that result in GD. Being able to link structural modifications of each mutation to the phenotypic variation of GD would enhance the understanding of the disease. The aim of this work is to understand the structural dynamics of wild type and mutant GCase.

A model of the complex of the enzyme GCase with its facilitator protein, Saposin-C (Sap-C) was generated using Protein-Protein docking (PPD). In this work, a knowledge-based docking protocol that considers experimental data of protein- protein binding has been carried out. Here, a reliable model of the enzyme GCase with its facilitator protein is presented and is consistent with the experimental data.

To understand the structural mechanism of function of the enzyme GCase, it was imperative to study its structural dynamics and conformational changes influenced by its interaction with other components including lipid bilayer, facilitator protein or substrate. Coarse-Grained MD (CG-MD) was employed to study lipid self-assembly and membrane insertion of the complex. Classical Atomistic MD (AT-MD) was used to study the dynamics of the interactions between different components of the simulation.

Furthermore, the results of ten different AT-MD simulations sampling 9 μs have been analysed. An activation method of GCase by Sap-C has been proposed, the change in conformation of GCase when its facilitator protein is present has been highlighted, through the stabilization of the loops at the entrance of the binding site. The differences in protein-protein binding when GCase is mutated have also been emphasised.

Finally, Anharmonic Conformational Analysis and Markov State Models have been used to build a kinetic model of the system. This model supports our activation mechanism hyphothesis.

IMPACT STATEMENT

In this thesis, I present our enhanced understanding of the structural mechanism of function of the enzyme Glucocerebrosidase and its role Gaucher's Disease. Gaucher's Disease (GD) is a rare recessive disorder, whose clinic manifestations ranges from death at early age due to neurological damage to asymptomatic adults, and is concomitant with other diseases such as Parkinson's Disease or Lewy Body Dementia. More than 250 mutations in the enzyme Glucocerebrosidase have been reported so far, that result in GD. However, no one has yet been able to link the clinical complexity of the disease to the structural implications of the mutations. Understanding the differences between the structural mechanism of action at atomistic level between the wild type and the mutants, would help us provide the crucial link between genotype and phenotype.

Secondly, I used a combination of data reduction methods including Anharmonic Conformational Analysis in conjunction with Markov State Modelling to construct a kinetic model and propose a potential mechanism of activation of Glucocerebrosidase in complex with its facilitator protein Sap-C. Structural dynamics of such a complex in a membrane environment has never been reported before.

Finally, the interactions at the protein-protein interface can potentially be exploited to manipulate the activation mechanism. This can have important ramification in the design of peptides that activate Glucoserebrosidase.

ACKNOWLEDGMENTS

Firstly, I wish to express my most sincere gratitude to my supervisor, Dr. Shozeb Haider, for his encouragement and guidance through the course of this work. Without his expertise and enthusiasm, the completion of this work would not have been possible.

I have furthermore to thank Dr. Barira Islam, for her support and help in overcoming some obstacles as well as her invaluable scientific and human advice.

I would also like to thank UCL School of Pharmacy for affording me this remarkable opportunity.

I am also grateful to all those colleagues that have accompanied, inspired and encouraged me through this long journey. Specifically, I would like to thank my friends from the University; Isa, Miriam and Teresa for the traffic of notes, the hours in the library, coffee on the grass and especially for the laugher and love, they are all such inspirational women and impeccable professionals. To my friend, Helen Dixon (UCD), for introducing me to the British academic universe and helping me in this intricate process with her advice, wisdom and goodness. To Christie (UCL/WIBR) for the hours outside the lab speaking about science, her help and inspiration. To the scientists and friends from the Office 113 at UCL SoP, especially Uche, for the grammar corrections of my abstracts including these acknowledgements, his smile every morning and the exchange of biscuits in the hunger evenings. To Era and Tian Su, because this PhD would not have been possible without the meals and coffee in the garden on the rare sunny days in London. Also, thanks to Yiannis for listening to my stories after work, his support and help.

This work is dedicated to my family. To my parents, Pedro and Marisol, for continually supporting me in so many aspects, and giving me my first microscope to play with. To my sister, Sol, and my brother, Pedro, for their faith in me and encouragement despite the distance, and to my nieces Emilia, Sol and Marta for illuminating me with their smile each step of this journey. Finally, this thesis is specially dedicated to my aunt ‘tia Lili’ who passed away during the writing of this thesis, we will not forget you, your laugh, your smell and your goodness, you were such a beautiful person.

To my parents

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: GLUCOCEREBROSIDASE- SAPOSIN- C: PROTEIN- PROTEIN MODEL.

CHAPTER 3: MOLECULAR DYNAMICS

CHAPTER 4: KINETIC MODEL

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

LIST OF ABBREVIATIONS

Gaucher's Disease (GD)

Glucocerebrosidase (GCase)

Glucosylceramide GluCer

Saposin-C (Sap-C)

Glucosphingolipids (GSL)

Sphingolipid Activator Proteins (SAP)

Lysosomal Storage Disorder (LSD)

Gene for the Lysosomal Enzyme Acid-β-glucosidase (GBA)

Parkinson Disease (PD)

Lewy Bodies Dementia (LBD)

Enzyme Replacement Therapy (ERT)

Food and Drug Administration (FDA)

Endoplasmic Reticulum (ER)

Mannose-6-Phosphate (M-6-P)

Lysosomal Integral Membrane Protein (LIMP-2)

Triose Phosphate Isomerase (TIM)

Activator ligand Isofagomine (IFG)

Sphingolipid Activator Proteins (SAPs)

Solvent Accessible Surface (SAS)

Complementarity Score (S)

Fast Fourier Transform (FFT)

Molecular Mechanics (MM)

Optimized Potentials for liquid Simulations (OPLS)

Ambiguous Interaction Restraints (AIRs)

Galactocerebrosidase (Galc)

Saposin A (Sap-A)

Molecular Dynamics (MD)

Quantum Mechanics (QM)

Energy Minimization (EM)

Periodic Boundary Conditions (PBC)

Coarse- Grained Molecular Dynamics (CG-MD)

Atomistic MD (AT-MD)

Alpha Carbon (Cα)

Root-mean square deviation (RMSD)

Root-mean square Fluctuation (RMSF)

Protein-Protein interactions (PPi)

Free Energy landscape (FES)

Anharmonic Conformational Analysis (ANCA)

Markov State Models (MSMs)

Kurtosis (κ)

Spatial Decorrelation (SD)

Temporal Decorrelation (TD)

FIGURES AND TABLES

CHAPTER 1: INTRODUCTION

CHAPTER 2: GLUCOCEREBROSIDASE-SAP-C PROTEIN-PROTEIN MODEL

CHAPTER 4: KINETIC STUDIES

CHAPTER 1:

INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1. General Introduction

Gaucher's Disease (GD) is a metabolism disorder with a variety of clinic manifestations that ranges from asymptomatic adults to early age death due to acute neurological damage.[1,2] It is a recessive disease caused by mutations in the gene that encodes the Glucocerebrosidase enzyme (GCase). Mutations yield a dysfunctional enzyme. GCase is a lysosomal hydrolase that cleaves glycolipids. Its malfunction causes glycolipid accumulation in lysosomes, which is responsible for the symptoms of GD. At least 250 mutations in GCase have been reported.[3] So far, there is no structural information that relates mutations in the enzyme with the different phenotypes of the disorder,[4] though few experimental techniques do permit the study of structure and function of individual amino acids mutations.[5]

In this study, we present a structural understanding of the molecular mechanism of function of Glucocererbrosidase in complex with its facilitator protein Saposin-C, including two commonly occurring disease phenotypes, using structural bioinformatics tools.

1.2. Introduction to membrane lipids degradation in Lysosomes and Lysosomal storage disorders.

Lysosomes are the major degradation organelles in eukaryotic cells. They are relatively small acidic compartments of spherical form containing a number of different enzymes necessary for degradation of biomolecules. Usually referred as stomach of the cell, these organelles represent a key point in cell homeostasis.[6,7] Components of eukaryotic membrane are broken down into their building blocks by lysosomal hydrolases and sent to the cytosol to be re-utilised, this being the principal mechanism of cell membrane turnover.[6-8]

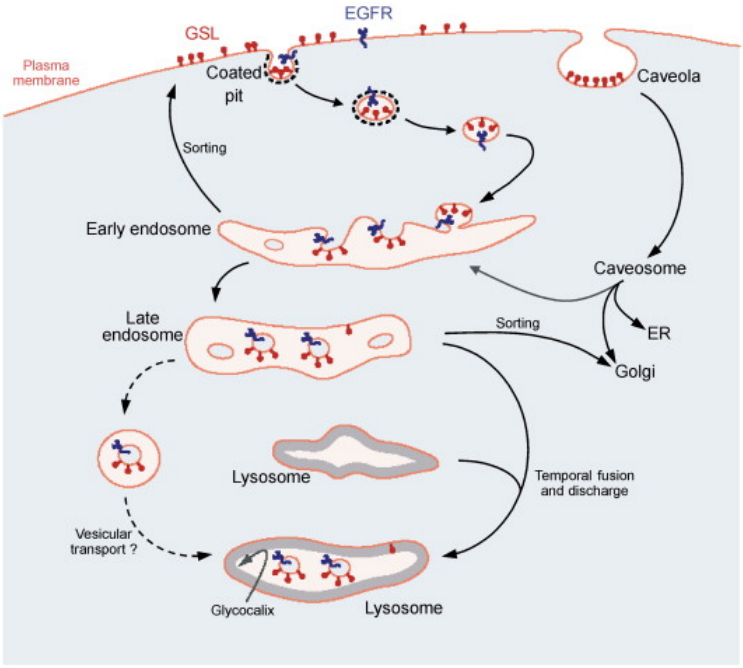

Figure 1.1*: Schematic illustration of the membrane component turnover. Membrane building blocks are endocytosed, carried to the lysosomal and degraded in the lysosomal lumen. Taken from reference 7.*

Glucosphingolipids (GSL) are ubiquitous components of eukaryotic cell membranes. They are complex glycolipids formed by a ceramide moiety and an oligosaccharide chain. These membrane components are degraded within the surface of intra-lysosomal membranes. The sequential action of different membrane hydrolases cleaves these components on their glycosidic bonds.[9] When the oligosaccharide chain of the glycolipids has less than four sugars, hydrolytic enzymes dissolved in the lysosol have difficulties to reach their substrate embedded in the membrane. In addition, they require the mediation of small protein cofactors that solubilise these complex lipids, thus making them accessible. These facilitator proteins are called Sphingolipid Activator Proteins (SAP).[7-10]

Mutations in genes encoding lysosomal hydrolases are responsible of a group of genetic disorders known as Lysosomal Storage Disorder (LSD).[10] As a consequence, undegraded macromolecules are accumulated in the lysosome causing severe clinical symptoms. Furthermore, accumulation of complex lipids such as GSL usually entails accumulation of other hydrophobic products via co- precipitation.[6,9,10]

This thesis focuses on the study of Gaucher´s Disease, a rare LSD produced by the deficiency or malfunction of the lysosomal hydrolase Glucocerebrosidase.

1.3. Gaucher ´s Disease: classification, pathophysiology and epidemiology.

Gaucher´s Disease (GD) is the most spread LSD known. It was first described by Philippe Gaucher, a medical student, in 1882. GD is a rare autosomal recessive disorder produced by mutations in the gene for the lysosomal enzyme acid-β-glucosidase or glucocerebrosidase (used hereafter). Mutations result in reduced hydrolytic activity, poor intracellular stability or diminished trafficking of the enzyme, causing accumulation of its principal substrate: glucosylceramide in the lysosomes.[1,3] The gene is located on chromosome 1 and it is composed of 11 exons, where single nucleotide substitutions, deletions, insertions and recombination with a highly homologous pseudo-gene (ps- GBA) located 16 kb downstream are observed.[3] Up to 250 mutations of this gene account for the wide range of symptoms seen in Gaucher´s disease.[3,11]

From lethality in early childhood to asymptomatic adults, GD shows a great variety in its symptomatology, including neurological involvement that ranges from eye movement disorder to severe neurologic degeneration. The disease has been classified into three types: non-neuropathic, acute and chronic.[10]

Type 1 GD or non-neuropathic (GD1): It is the most common form of GD and is normally diagnosed in late childhood, or early adulthood although some patients remain asymptomatic throughout their entire life. It is characterised by hepatosplenomegalia, thrombocytopenia which causes spontaneous bleeding in some cases, anaemia, pain crisis, bone involvement, including pathological fractures, osteoporosis or deformities, skin involvement, and in rare cases lung involvement, such dyspnoea, or pulmonary hypertension, and heart involvement. [1,2,11]

Type 2 GD or acute neuropathic (GD2): This form is characterised by a progressive neurological degeneration causing individuals to die by age 2 years. The symptoms include limb rigidity, spasticity or eye movement failure. Patients with this type of GD typically present strong skin involvement due to modification in the ratio glucosylceramide/ceramide in the stratum corneum. GD2 may also result in hydrops fetalis in some cases. [1,2,11]

Type 3 GD or chronic neuropathic (GD3): The course of this type is more similar to that of GD1 with mild neurological involvement that range from eye movement defects such as saccade initiation failure or strabismus to learning disabilities, developmental delays, autism disorder and in some cases dementia, convulsions or ataxia. .[1,2,11]

Although GD is a pan- ethnic disease it has been extensively reported to have a greater incidence in Ashkenazi Jews, with roughly 1 case in 800 live-births for this population against the approximately 1:50000 in general population. Among the different forms of GD, type 1 is by far the most common with roughly 1:40000 cases, whereas the other two forms occur in 1 individual in every 100000. [1,2]

The pathophysiology of GD still remains unclear. Although glucocerebrosidase and its main substrates glucosylceramide (GluCer) and glucosylsphingosine are present in every cell, the complex lipids are especially abundant in red and white cell membranes. Thus, macrophages would be the most affected cells, incapable of metabolising the complex glycolipid after the blood cells turnover.[1,11] Consequently, macrophages, fattened with GluCer, would be accountable for the extensive organomegalia. As osteoclasts also belong to the mononuclear phagocyte lineage, could also be affected producing the skeletal disease. Therefore macrophage has become the centre of study to understand the mechanism of GD and to develop a treatment. However, this seems to be unsatisfactory to explain the complex symptomatology.[12] More recently, deregulation of other immune cells, apart from the macrophage, has been reported in clinical studies.[13] Furthermore, it has been demonstrated that the bone involvement could be more related to an inhibitory action by complex glycosphingolipids of osteoblastogenesis. Finally, extramedullary haematopoiesis was identified that may be producing the organomegalia. These latest studies have given new perspectives over the study of this complex disease.[12,13]

Apart from the complex symptomatology and the phenotypical variation, some studies have shown how GD may predispose to the development of other disorders, or at least that there is certain concomitance with other diseases that deserve to be highlighted.[3] A well-studied association has been between GD and Parkinson Disease (PD).[2] Mutations in GCase have been found in samples taken from PD patients in higher proportion than in the control group.[14] On the other hand, some cases of GD have been

reported to show Parkinsonian manifestation in adulthood.[15] Another proposed association has been between GD and Dementia with Lewy Bodies (DLB), similar to that observed with PD. Some cases with confirmed DLB have been found to carry mutations in GCase; moreover GD cases have been reported to develop Lewy bodies.[16] Finally, GD has also been associated with some types of malignancies including multiple myeloma and haematological cancers.[1,2,11]

Few strategies have been tested for the treatment of GD. Administration of exogenous enzyme or enzyme replacement therapy (ERT), easily recognisable by macrophages is still the treatment of choice in most of the cases. This strategy has shown to be effective reducing organomegalia, anemia and cytopenia. However, it has the disadvantage of not passing the blood-brain barrier, being unable to ameliorate bone and lung disease, as well as its price and its administration route: parenteral.[1] Substrate reduction by partial inhibition of glucosylceramide synthase, first enzyme in the route of the biosynthesis of glucosphingolipids has been another strategy for the treatment of GD.[17,18] This is the case of N-butyl-deoxynojirimycin (miglustat), which was the first of these agents to be approved by the Food and Drug Administration (FDA) as valid treatment for patients who presented any contraindication to ERT.[19] Oral pharmacological chaperones, or agents that ease protein folding, and gene therapy are being the latest areas to be explored for the treatment of this disease.[1,11,13]

1.4. Glucocerebrosidase: function and structure.

Glucocerebrosidase also called acid-β-glucosidase (GCase), is a lysosomal enzyme composed of 497 amino acids. GCase is synthesised in the rough endoplasmic reticulum (ER) as a pre-enzyme with 19 amino acids leader polypeptide in its N-terminus.[4] The pre-enzyme loses this sequence as part of the complex post-translational events that occur prior to its traffic to the lysosome.[4,20,21] Unlike other lysosomal hydrolases, GCase is not marked with mannose-6-phosphate (M-6-P) for its traffic to the lysosome, but its association with Lysosomal Integral Membrane Protein (LIMP-2) assists in its entrance in the organelle.[20,21]

GCase is a hydrolytic enzyme that cleaves its main substrate Glucosylceramide (GluCer), into glucose and ceramide. Although GluCer is its main substrate, the enzyme also breaks down the des-acyled form of this glycolipid - glucosylsphingosine.[4] The lipidic tails of both glycolipids are embedded in the intra-lysosomal membrane such that both substrates lay inaccessible. This makes GCase need the assistance of a second facilitator protein Saposin-C (Sap-C) to anchor to the membrane. Sap-C belongs to the Sphingolipid Activator Proteins. Unlike other members, it does not only mediate the contact of the GCase with its natural ligands, but it is also known to be able to stimulate the enzyme activity directly.[22,23] Recent investigations have revealed that both GCase and Sap-C associate in the membrane.[4,7,23]

GCase is a globular protein. As it is shown in Figure 1.3, this enzyme is composed of three different domains: Domain I (residues 1-27 and 383-414) is a small three stranded anti-parallel β-sheet, Domain II (residues 30-75 and 431-497) is an independent 8 stranded β-barrel resembling an immunoglobulin domain and Domain III (residues 76-381 and 416-430) is a $(\alpha/\beta)_8$ TriosePhosphate Isomerase (TIM) Barrel, containing the active site. Domains I and III interact tightly and are linked by one of the loops at the entrance of the binding site. Domains II and III are separated by a long loop that acts as a hinge. Structural folds similar to Domains II and III can be found in other hydrolases such as α- galactosidase. [4,25]

In GCase, residues E340 and E235 have been identified as the catalytic residues within the active site.[24] Residue E340 is the catalytic nucleophile and E235 the acid-base

residue. In the first step, E235 would supply a proton to the ceramide group.[4,24,25] As illustrated in Figure 1.2, a nucleophilic attack is initiated by the oxygen atom of the residue E340 on the anomeric carbon. An intermediate adduct is formed, following by its hydrolysis.

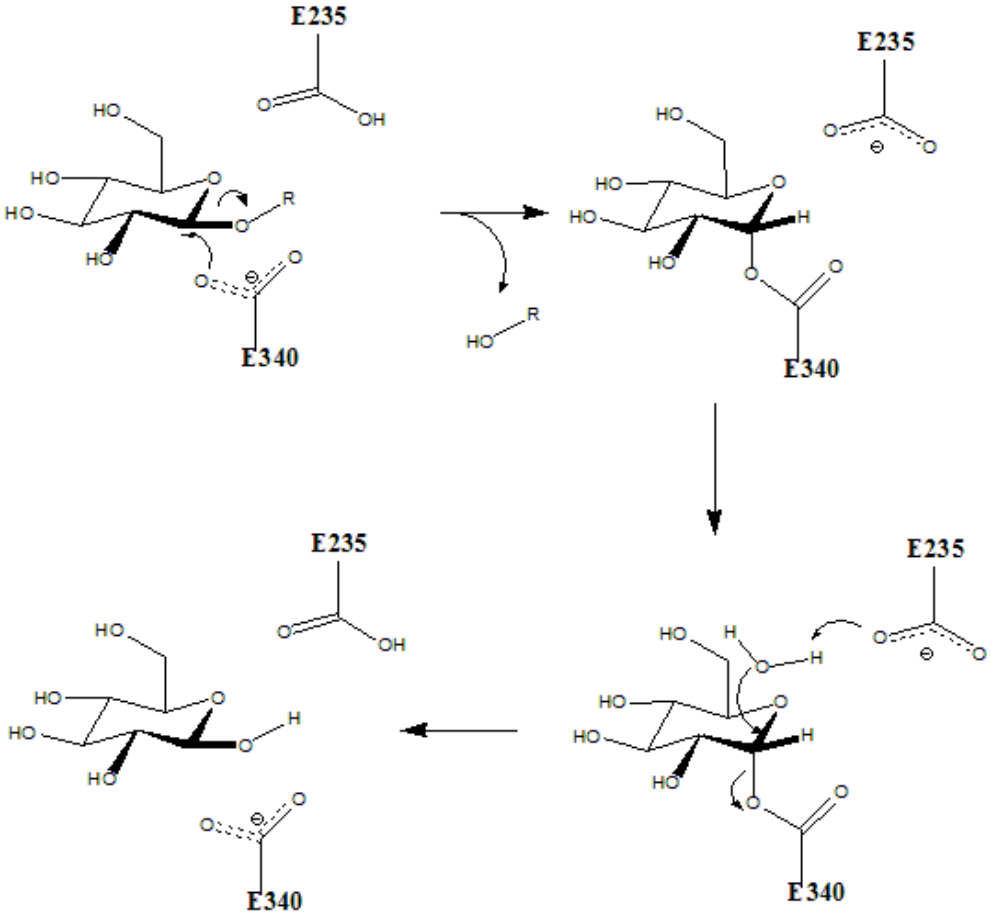

Figure 1.2: *Catalytic mechanism of the enzyme GCase. The nucleophilic attack by E340 results in the formation of an adduct and release of the ceramide group (OH-R). Acid base residue E235 supplies a proton to the ceramide group. The adduct is hydrolysed and a molecule of glucose is realised.*

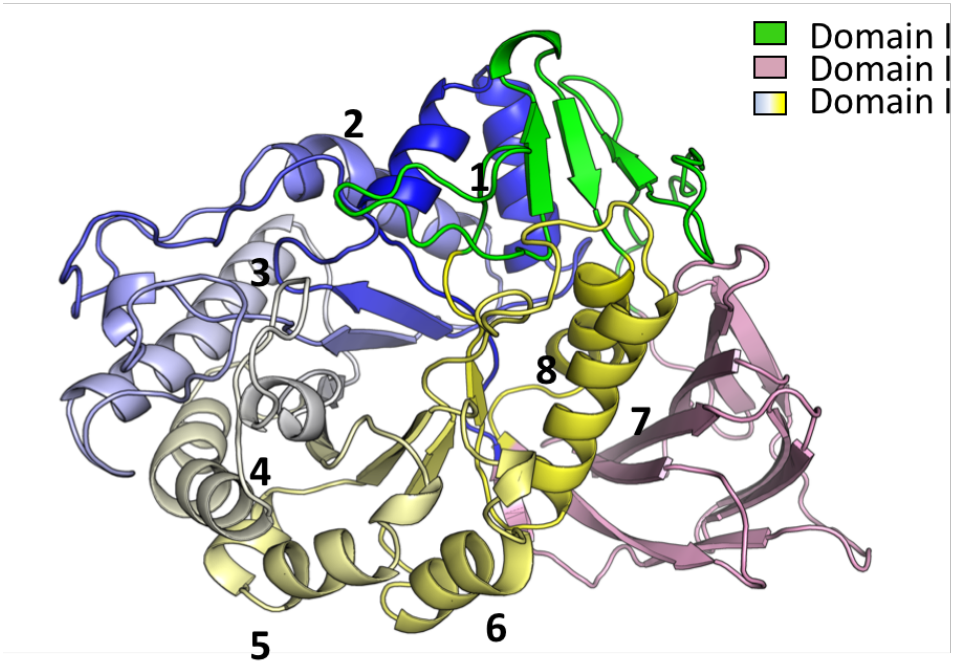

Figure 1.3: *Structural arrangement of Glucocerebrosidase. The enzyme consists of 3 different domains: Domain I is a three-stranded anti-parallel β-sheet (green), Domain II is an 8 stranded β- barrel (pink) and Domain III is the TIM barrel (yellow to blue). The numbers indicate the name of helices in the TIM barrel.*

The active site lies in a cavity formed at the centre of the TIM barrel motif. Residues R120, D127, F128, W179, N234, Y244, F246, Y313, C342, S345, W381, N396, F397 and V398, constitute the glucose-moiety binding region, and residues E235 and E340 are the catalytic residues (Fig. 1.4).[4,24,25] Some of these aromatic residues may play an important role in molecular recognition of the substrate.[26] Other residues, inside the pocket, create a hydrogen-bonding network that holds the substrate in the correct position for hydrolysis. Some of these residues do not change their conformation upon GluCer binding, whereas others are flexible. Particularly, Y313, N396 and F397 show high thermal B-factors in crystal structures.[27] It is not clear what parts of the active site are implied in holding the ceramide tails. It is assumed that they are embedded in the membrane during the catalysis.[4,24]

Five loops (Loop-1 (residues 311-319), Loop-2 (residues 345-349), Loop-3 (residues 394-399), Loop-4 (residues 237-248) and Loop-5 (residues 283-288)) at the entrance of the active site are believed to rearrange in different conformations allowing and blocking the entrance and holding the substrate in the active site (Fig. 1.5)[4]. Different conformation of these loops has been reported in the different available crystal structures of the GCase. Especially important are the changes in Loop 1, Loop2 and Loop 3, that determine the accessibility to the binding site.[4,24,27]

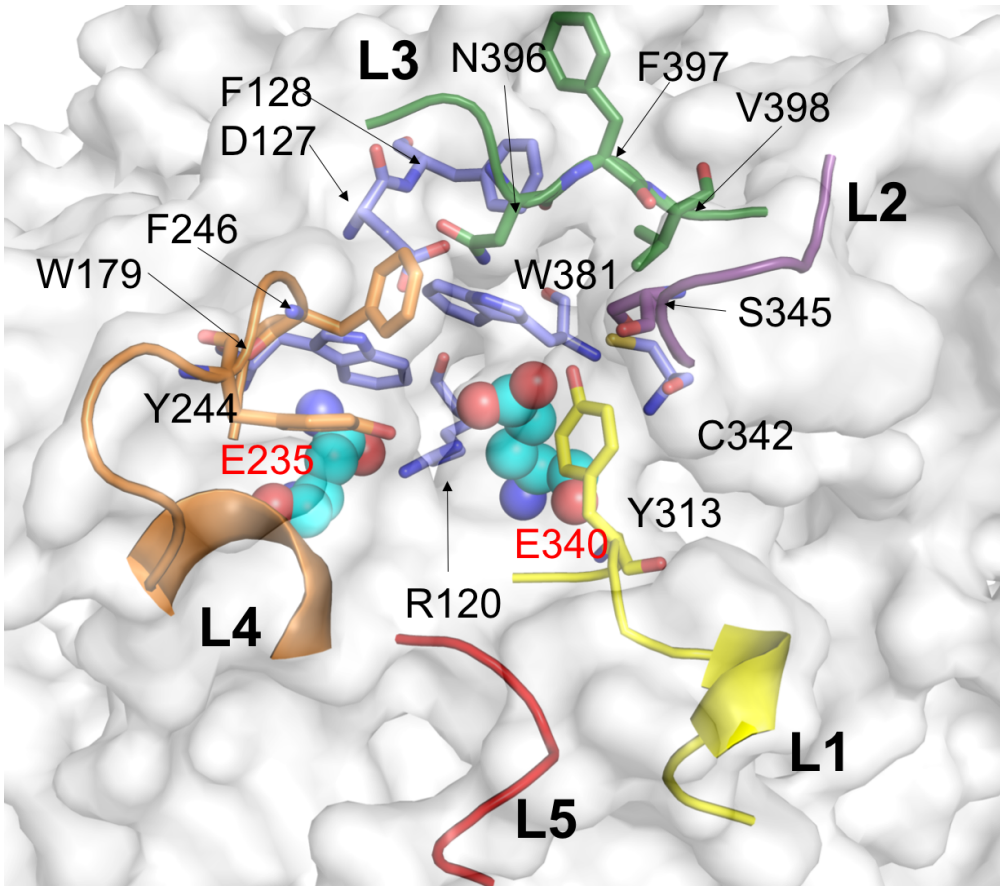

Figure 1.4: *Residues in the binding site. Five loops at the entrance of the binding site have been labelled: Loop-1 in yellow, Loop-2 in purple, Loop-3 in green, Loop-4 in orange and Loop-5 in red. Catalytic residues (E235 and E340) have been coloured in cyan, represented as spheres and labelled in red. The rest of residues of the binding site are in dark blue.*

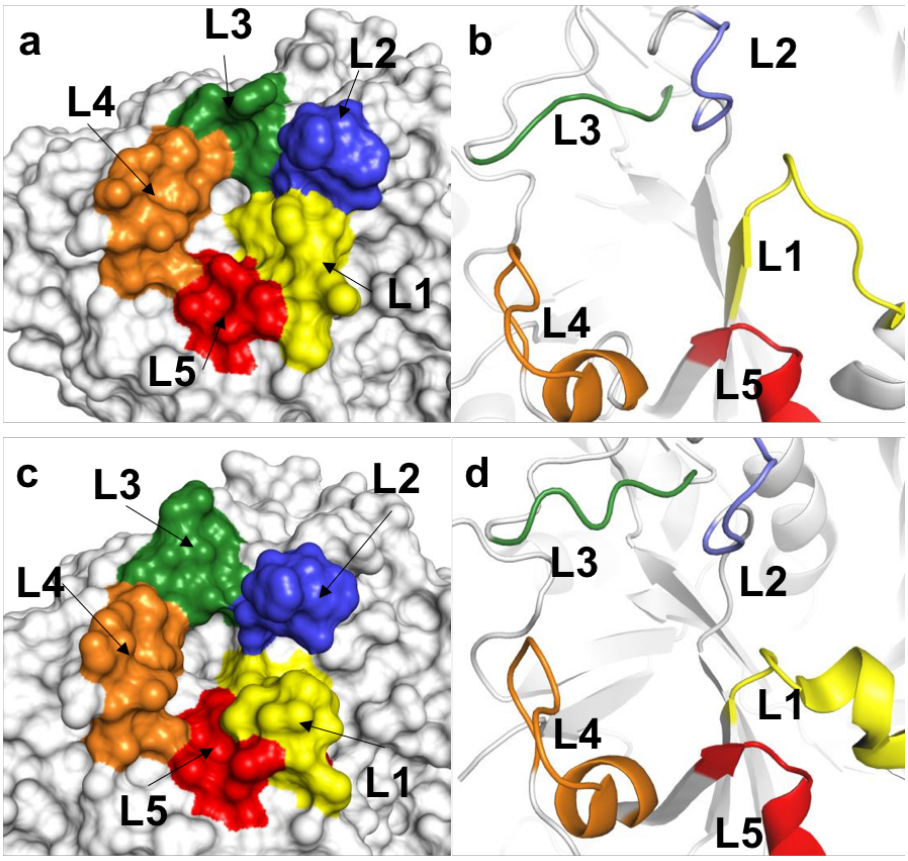

Figure 1.5: *Conformational changes in the active site loops. Rearrangement of loops at the entrance of the active site in (**a** and **b**) inactivated GCase (PDB id 1OGS), where the L1 is in extended conformation and (**c** and **d**) activated GCase (PDB id 2NSX), where L1 adopts helical conformation. Individual loops have been labeled L1-L5.*

Two different configurations of the Loop-1 (residues 311-319) have been reported: extended and helical, are shown in Figure 1.6. Residue D315 in the middle of the loop leans towards loop 2 in the extended conformation, establishing hydrogen bond with G244. Yet, in the helical conformation the residue seems to tend towards residue N370, to form a water-mediated hydrogen bond with the N370 and a salt-bridge with R285. Residue Y313 changes from establishing hydrogen-bonded interaction with E235 to E340 when the loop conformation shifts from extended to helical conformation.[4] Residue W312 also changes hydrogen bonding partner with a change in conformation from R285 in extended to C342 in the helical conformation. Activator ligand Isofagomine (IFG) has been reported to be better held inside the pocket when Loop-1

is in helical conformation.[27] Crystal structure of mutant N370S shows extended conformation.[28] Molecular docking studies demonstrated that extended Loop-1 clashes with the binding site. Moreover, GlcCer could not be properly positioned within the binding site when the Loop-1 is in the extended conformation.[4,27]

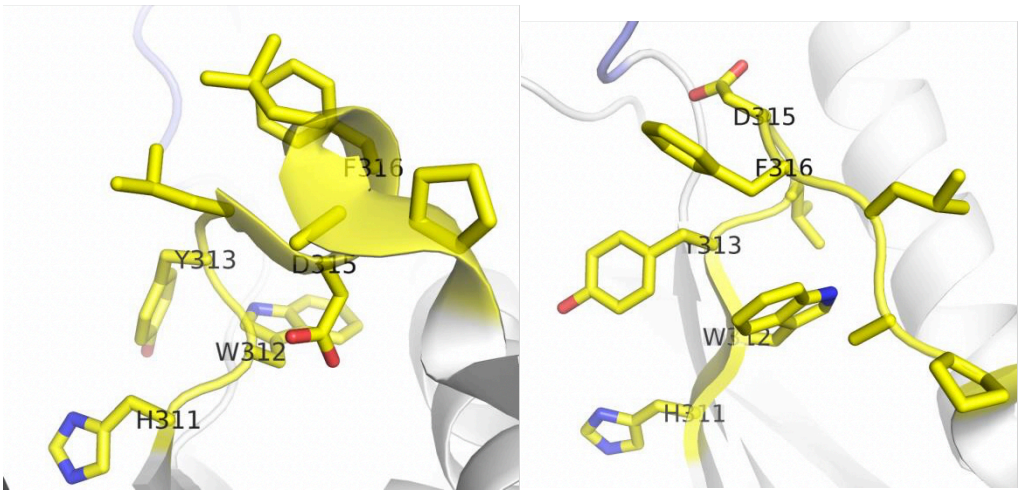

Figure 1.6: *Conformations of Loop-1 (yellow) are illustrated in* ***(a)*** *extended and* ***(b)*** *helical conformations. The movements in the side chains of the residues W312, Y313 and D315 are prominent. Loop-1 is in helical conformation when GCase is bound to an activator ligand in the active site.*

Three residues, (R395, N396, F397) from Loop-3 play important roles in the availability of the active site. In the apo structures when Loop-1 in extended conformation (where active site is not available) R395 is pointing towards Y313 and makes hydrogen bonds with the catalytic residues (E235 and E340). The side chain of F397 is oriented towards the centre, while the side chain of N396 is directed away from the active site. The orientation of F397 side chain and hydrogen bond network of R395 blocks the entrance of the active site. In the helical conformation of Loop-1, the side chains of R395 and F397 point out of the active site, while the side chain of N396 points inwards (Fig. 1.7). This creates a cavity that can accommodate the substrate.[4,27]

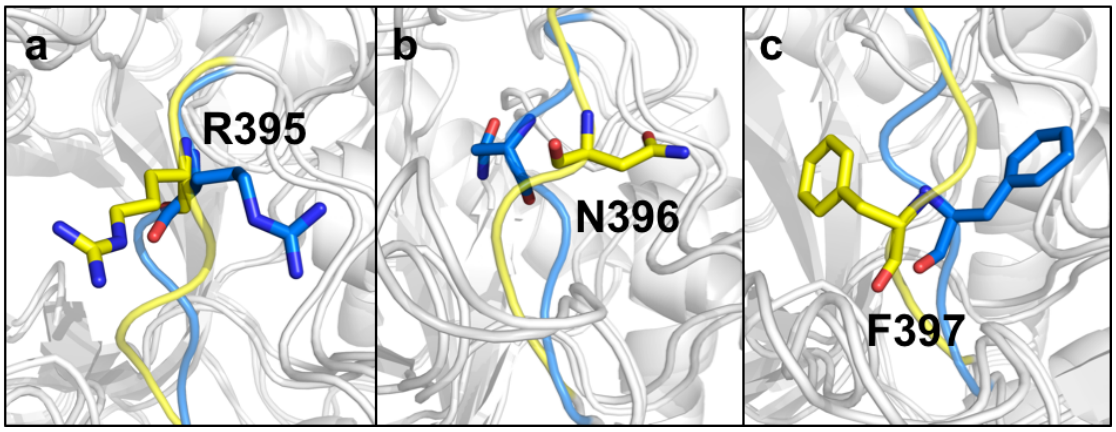

Figure 1.7: *Different conformations of the residues* ***(a)*** *R395,* ***(b)*** *N396 and* ***(c)*** *F397 in Loop-3 at the entrance of the active site. Activated GCase (pdb code: 2NSX) has been coloured in blue, whereas Apo-GCase (pdb code: 1OGS) has been coloured in yellow. Please note that R395 and F397 are pointed towards inside of the binding pocket when in inactivated conformation, thus impeding the entrance of the ligand in the active site. N396 is pointing towards inside of the active site when in activated conformation, probably having a role holding the ligand through electrostatic interactions in the active site.*

As the lysosomal enzyme requires sulphate or phosphate ions in the crystallization media, it has been proposed that some of the ions that remain in the crystal structure may coincide with some of the sites of lipid membrane binding. Two sites in particular have been identified as possible places of association to the membrane, one among the residues S12, R44, R353, S356, R357 and D358, and another formed by the residues K79, W228, R277 and H306. [29]

It has been reported that although glycosidation did not affect to the catalytic activity or activation via Sap-C, it was nevertheless essential for the formation of the active enzyme. Five glycosidation sites have been identified at Asparagines N19, N59, N146, N270 and N462.[30]

1.4.1. Mutations in Glucocerebrosidase

Although approximately 250 mutations of the enzyme have been reported so far, only one, N370S, is responsible for more than the 70 per cent of the cases of GD type 1.[31,32] The mutations D409H and D409V are important mutations in GD type 3,[33] L444P have been associated with neurological symptoms[34,35] and E326K does not produce GD but has been extensively reported to be found in patients with PD[16,36-37].

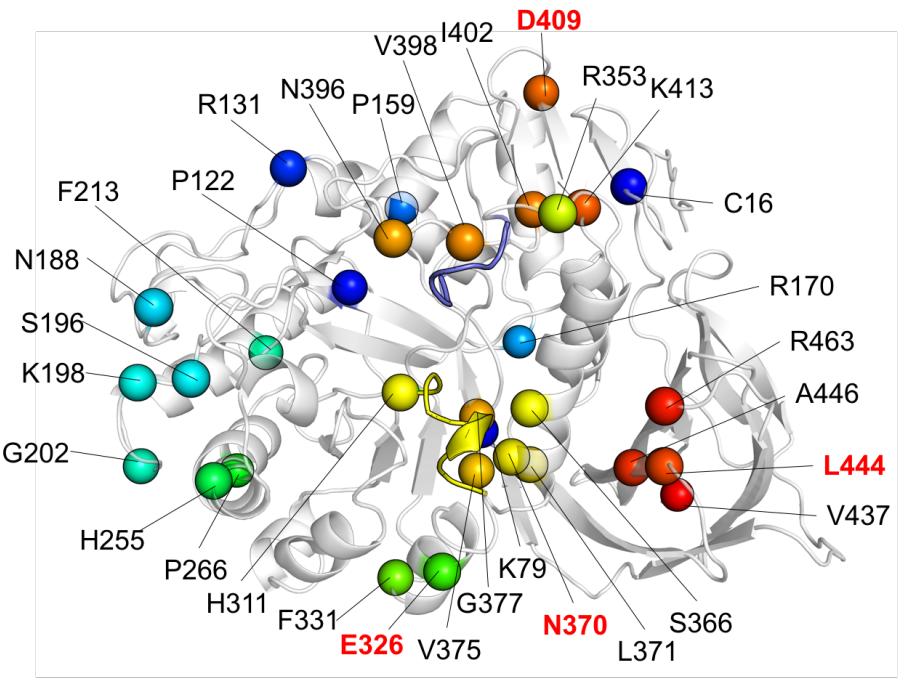

Figure 1.8: *GCase residues that cause Gaucher's Disease when mutated. Reported GCase residues mutated in Gaucher's disease have been represented as spheres, Loop-1 and Loop-2 at the entrance of the binding site have been coloured in yellow and purple respectively. Residues N370 (70 % of the cases of GD1), L444 (Neurological symptoms), D409 (GD3) and E326 (PD) have been labelled in red.*

1.5. Saposin-C: Function and structure

Saposin-C (Sap-C) is a small intra-lysosomal membrane protein. It is composed of 78 amino acids and it is about 10 kDa. It belongs to the Sphingolipid Activator Proteins (SAPs), which are non-enzymatic membrane proteins that mediate the association of proteins with their lipid substrate in the intra-lysosomal membrane.[7] Sap-C derives from the proteolytic cleavage of prosaposin in four highly homologous proteins Sap A-D. Despite their resemblance and high homology all four proteins are specific for different lysosomal hydrolases, and its absence or malfunction produces different LSD.[7,38]

The structure of Sap-C consists of four amphipathic α-helices forming a compact hydrophobic core and hydrophilic side chains on the surface (Fig. 1.9). Three di-sulphide bridges tightly bind the helices together. Apart from this, another conformation of the protein has been detected under detergent conditions. An open conformation of the saposin, in which the hydrophobic core is left exposed, has been identified under solvation with SDS. Sap-C has been crystalized as a monomer, but under low pH and detergent conditions, it has been reported to self-associate in dimers and trimers. Some studies have suggested that these associations may be relevant for the mechanism of action of these proteins.[37]

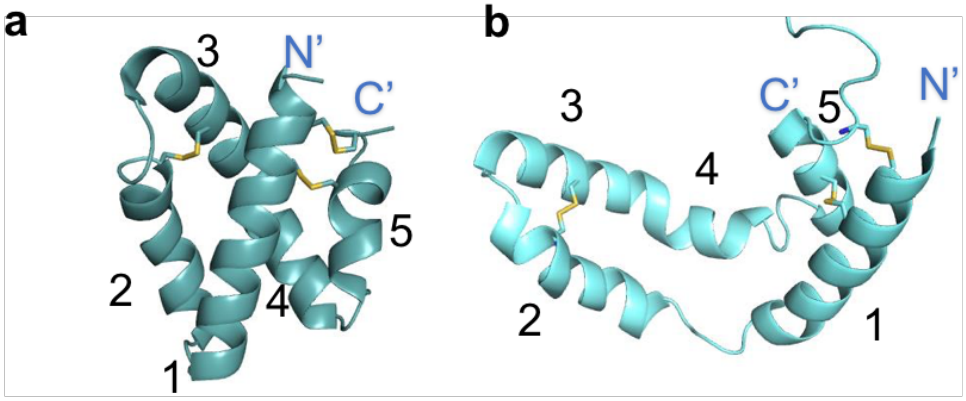

Figure 1.9: *Conformations of Sap-C.* ***(a)*** *closed and* ***(b)*** *open conformation. Numbers indicate the name of the helices. Disulphide bridges have been illustrated as sticks.*

The mechanism by which Sap-C destabilizes intra-lysosomal vesicles to make GluCer

accessible to its hydrolytic enzyme is not well known. The binding to the membrane is known to occur in a reversible pH dependent manner. After the negative charge on the surface of Sap-C is neutralized owing to an acidic pH, the protein is ready to bind to the phospholipids of membrane to carry out its action. It is known that Sap-C not only mediates the interaction of GCase with its natural substrate but also directly induces a conformational change in the hydrolase that allows it to complete its catalytic action.[22] It has also been demonstrated that Sap-C may protect GCase against the proteolytic action of lysosomal proteases.[39]

As a consequence, mutations in the GCase that affect the association to Sap-C would result in not only diminished activity, but also in more vulnerability of the enzyme to its early digestion in the lysosome, which would produce Gaucher's Disease. Furthermore, the mutations in the gene of prosaposin, that lead to the malfunction or absence of Sap-C in the lysosomal compartment were reported to be cause of a juvenile form of Gaucher's disease. [40,41]

1.6. Interaction of Glucocerebrosidase with Sap-C.

The area of interaction of GCase and Sap-C has not been accurately established. Experimental studies demonstrated that the GCase mutant N370S had decreased capacity of binding to Sap-C and to the phospholipid membrane. Since then, it has been postulated that the binding site of Sap-C in GCase should lie in the vicinity of N370.[42]

On the other hand, other experimental studies have focused on finding what domains of Sap-C that interact with GCase.[43,44] A study, carried out through competition assays of synthetic lipids, revealed two domains capable of binding and activating GCase. They included Domain 1: residues 6-34, binding site: 6-27 and activation site: 27-34 and Domain 2: residues 41-60, binding site: 45-60 and activation site: 41-49. Domain 2 in that case was reported to bind GCase by at least one order of magnitude more strongly.[45] Another study, conducted with chimeric saposins, determined the activator region of Sap-C to be located between residues 47-62.[43]

1.7. Molecular modelling and structural bioinformatics techniques

Molecular modelling comprises of all those computational or theoretical methods that can be used to model or emulate the behaviour of biomolecular systems *in silico*. These methods include techniques to predict the binding mode of two interacting biomolecules (molecular docking), methods to predict the evolution in time of a biomolecular system, such as molecular dynamics simulations (MD) or methods to model the kinetics of the systems. Thereby, molecular modelling provides a valuable strategy to understand a biological system at a detail, which is impossible to obtain by other experimental techniques. Computational and theoretical methods are especially suitable for the study of the implications of different mutations in a protein, when structural details need to be known. The computational techniques used to carry out this research are going to be introduced in detail in successive chapters.

1.8. Previous Studies

A computational docking model for the interaction of GCase with Sap-C was proposed by Atrian et al. in 2008.[46] The model relies on both structural and evolutionary information of the pair of proteins. Through the study of correlated mutations in seven pairs of homologous proteins of different species, they identified residues, which are important for the protein-protein interaction. This information was then used to limit the search of docking poses. This investigation proposed that the binding site of Sap-C with GCase is located between the helices 6 and 7 of Domain III and Domain II, and include the residues: L314, L317, A318, W348, D358, Q362, H365, S366, T369, N370, L372, Y373, K441, D443-D445, R463, S464 and Y487. The interacting residues of Sap-C proposed to form part of the interaction were: E9, D20, E25, S56, S57, L59, S60, L62- S67, E69, L70 and M74. The protein-protein docking was carried out between an inactive structure of GCase and an open and closed conformation structure of Sap-C.

While the proposed model is a good starting point to understand GCase and Sap-C protein-protein interactions, it has two fold limitations. Firstly, it is unable to account for experimental data, as the residues they proposed to form part of the binding site do not coincide with the experimental studies mentioned[44,43]. Secondly, there is no structural information regarding the anchoring of the complex with the membrane.

Few attempts have been made to explain the different effects of mutants via Molecular Dynamics simulations. In 2007, Zubrzycki et al. aimed to understand the behaviour of two mutants, namely: L444P and L444R.[47] They carried out 1 ns molecular dynamics in explicit solvent. They also performed blind ligand docking to identify the most metabolically important residues. The ligands used for the docking experiments were: substrate (GluCer), an inhibitor (conduritol-β- epoxide) and the product (glucose). The conclusions drawn from this study were, although L444P and L444R result in the same phenotype of Gaucher´s disease (type 3), they might be a result of different structural consequences. While L444R conceals the hydrophobic core of domain II due to steric occlusion, L444P was shown to lower the flexibility of loop 2 at the entrance of the binding site. Residues D443 and D445 were demonstrated to be somehow implicated in the binding of the ligands in the active site.

In 2010, Offman et al. studied the dynamics of the N370S and other important mutants (F213I, D409H, L444P and R496H).[48] A 10 ns molecular dynamics simulation study was intended to provide an explanation to the reduced activity of the mutants. An active form of the enzyme was used as a starting structure. In the N370S simulation, a change in hydrogen bonding pattern in the active site was observed when it was compared to the wild type simulation. A change in contact pattern made the binding site smaller and less accessible for the ligand. In addition, a change in the conformation of the loop 1 in the vicinity of the binding site was also reported. The study also found that the said changes were reverted upon the binding of a pharmacological chaperone in the mutated enzyme.

The studies reported by Zubrzycki et al. and Offman et al. explained structural interactions in selected mutants. However, they lacked information on the interactions with Sap-C, membrane anchoring and the influence of membrane lipids/substrate on the structure of GCase/Sap-C complex. Finally, the time scales of the molecular dynamics simulations were too short to produce meaningful results.

1.9. Aims and objective of this thesis

In order to improve our understanding of the structural role of GCase in Gauchers disease, this thesis aims to:

1. Construct a knowledge-based protein-protein model of GCase-Sap-C complex
2. Identify how the complex binds to the membrane
3. Understand dynamics of interactions at the GCase-Sap-C interface
4. Investigate the influence of Sap-C on the activation mechanism of GCase
5. Compare the differences between wild-type and mutant GCase in complex with Sap-C.

1.10. Preliminary conclusions and goals achieved

In this chapter, I have explored the biological importance of the enzyme Glucocerebrosidase whose lack of activity leads to the most common Lysososmal Storage Disorder - Gauchers Disease. Understanding the structural activation mechanism of this lysosomal enzyme is the cornerstone for unravelling the complex phenotypic profile of Gauchers disease.

In this thesis, a combination of computational tools has been employed to elucidate the activation mechanism of the enzyme Glucocerebrosidase by its facilitator protein Saposin-C. Firstly, a knowledge-based docking protocol that considers experimental data of protein- protein binding has been used to generate the Glucocerebrosidase-Saposin-C complex. Next, a multiscale molecular dynamics simulations have been employed to study lipid self-assembly, membrane insertion and dynamics of the interactions between different components of the complex. Based on a total sampling of 9 μs, we propose a model that explains the structural activation mechanism of the enzyme GCase facilitated by its activator protein Saposin-C. Conformational changes in the loops at the entrance of the binding site are stabilized by direct interactions with Saposin-C. A loss of interactions with Saposin-C, result in destabilization of the complex and thus explaining the structural basis of pathophysiology arising in N370S and L444P Glucocerebrosidase mutants.

CHAPTER 2:

GLUCOCEREBROSIDASE- SAPOSIN- C: PROTEIN- PROTEIN MODEL.

CHAPTER 2: GLUCOCEREBROSIDASE- SAPOSIN- C: PROTEIN-PROTEIN MODEL.

2.1. Introduction

Understanding the interaction of GCase with its facilitator protein, Sap-C, was the first step in our research towards a comprehensive explanation of the structural mechanism of function and implications of the mutations causing Gaucher's disease. The existence of high resolution 3D structures of individual components including GCase and Sap-C, as well as the advanced docking algorithms encouraged us to generate computational models of the GCase-Sap-C complex.

Molecular recognition occurs due to the formation of specific attractive interactions at the protein-protein interface. Computationally predicting those highly specific interactions has been called "the docking problem", and different algorithms to solve it have been probed since the eighties.[49,50] When two interacting partners are as big as two proteins, determination of the binding mode becomes very challenging due to the multiple degrees of freedom. Consequently, such algorithms normally neglect the movement of the side-chains and protein domains. The proteins are treated as relatively rigid bodies and limit the search to six degrees of freedom. On the other hand, the site of interaction in the protein surface is not as obvious as it is when it interacts with its substrate, making the search even more difficult. The availability of experimental data dramatically accelerates the search and increases the chance of obtaining the correct binding mode.[51,52,53]

Solving the docking problem requires three basic components, namely: a mathematical model that effectively represents the system, usually based on geometrical features of the protein surface; an efficient search algorithm able to explore the conformational space at a reasonable speed and a scoring function, that scores and ranges the different conformations relying on energy terms and/or shape complementarily, that should ideally be able to distinguish the native conformations and reject the non-native ones. Moreover, in protein-protein docking it is also necessary to implement a subsequent energy-relaxation method, since the proteins are treated as rigid bodies, the solutions are prone to have steric clashes.[51]

Different types of the components mentioned above and their different combination define the variety of docking methods currently available. In this project, we have used two docking programs: Hex and Haddock, the details of which are explained below.

2.1.1. HEX

Hex is a protein-protein docking method in which protein surfaces are depicted as skins constructed using Fourier expansions of spherical polar coordinates of N orders that represent protein surface shape, charge, density and surface electrostatic properties.[55,57] Unlike geometric search of protein shape complementarity, Fourier Correlation methods have the advantage of computing the level of overlap between two systems, represented as Cartesian grids, very rapidly and automatically penalising steric clashes between them, thus accelerating the search. [54] Furthermore, this method allows to perform a low-quality search by decreasing the number of expansion orders (polynomial powers). Lower order expansions decrease the quality of the representations which facilitates a quick preliminary analysis of the less relevant conformations. The reduction of the expansion orders also provides some degree of softness in the representations which allows to incorporate certain protein flexibility. [51,55,56]

Protein shapes are defined as Gaussian density functions. The volume between the solvent accessible surface (SAS) of the proteins and the van der Waals surface of the protein atoms is the so-called skin. So that, the van der Waals surface of the protein would be the interior skin and can be calculated as the sum of the relative contribution of each atom to the density function and can be represented in a 3D grid. The SAS bounds the limits of the exterior skin which is defined by making a test sphere roll over the van der Waals surface. Interior and exterior skins are represented as density functions τ ($\underline{r}$) and σ ($\underline{r}$) respectively, which equals to one inside the skin and zero anywhere else, as shown in equation 2.1 and 2.2.[53-55]

$$\sigma(\underline{r}) = \begin{cases} 1; & r \in outer\ skin \\ 0; & \text{Otherwise} \end{cases} \qquad \tau(\underline{r}) = \begin{cases} 1; & r \in surface\ atom \\ 0; & \text{Otherwise} \end{cases} \qquad (2.1, 2.2)$$

Density functions can be approximated as expansions of various orders N. Increasing N results in an increase in shape resolution, but also in computational time. Medium orders are recommended to conduct a low-quality search that allows rejecting the least relevant conformations. In subsequent steps, higher N may be used for a softer analysis of the best solutions.[53-55]

The idea behind the use of the skin representation of proteins is to maximize the overlap between the interior skin of one protein and the exterior skin of its docking partner. The shape complementarity score (S) accounts for the volume of solvent displaced upon association and penalizes the steric clashes produced by the interior skin overlap of both proteins:[55]

$$S = \int (\ \sigma_A(\underline{r}_A)\ \tau_B(\underline{r}_B) + \ \tau_A(\underline{r}_A)\ \sigma_B(\underline{r}_B))\ dV - Q\ \int \tau_A\ \tau_B dV \qquad (2.3)$$

where the first term refers to the volume of solvent displaced and can be used as an approximation to the hydrophobic free energy association, and the second term accounts for the interior-interior skin overlap, multiplied by a penalty factor Q=11.

The search method of the conformational space in Hex differs considerably from that in former Fast Fourier Transform (FFT) algorithms. The search is carried out in gradual rotations more than in translational searches, which simplify the transformation of coefficients of the spherical polar parameterization. The space search is covered in six degrees of freedom, five Euler angles, four rotational angles (two for each protein, β1, γ1 and β2, γ2) and one axial (α) and one translational degree (Fig. 2.1). Each protein rotates around its own centroid in gradual increments of the rotational angles generated from icosahedral tessellations of the sphere; the distance between centroids also varies in translation operations. A common coordinate system for both proteins is assumed. [53,55]

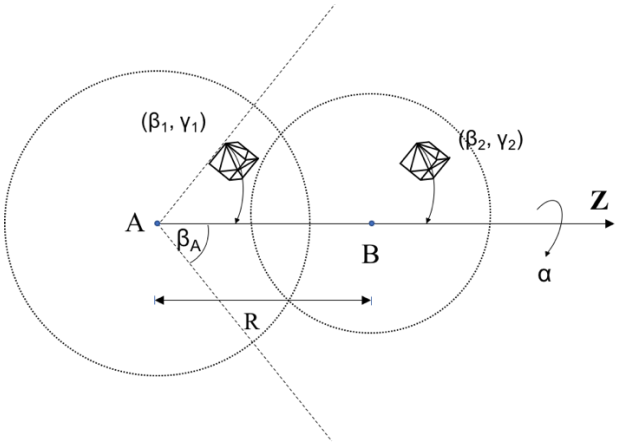

Figure 2.1: *Representation of the docking search using icosahedral tessellation. A and B represent the coordinate origin of the two proteins. Where β and γ are the molecular rotational increments generated by tessellations for each protein. α indicates the twist angle.*

Hex also offers different options for docking post-processing. The simplest option is to count steric clashes between not bonded atoms through a bump counter. The program also performs Molecular Mechanics (MM) refinement based on hydrogen bond (12-10) and soft Lennard-Jones (12-6) potentials, and it is an adaptation of Optimized Potentials for liquid Simulations (OPLS)[58] force field.[55]

The final docking score in Hex is the sum of the shape-based correlation energies, which is a combination of S (complementary score) and electrostatic energy, and the OPLS energy at the minimized orientation, as shown in equation 2.4. [53]

$$E_{TOTAL}(R, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2) = E_{SHAPE} + E_{OPLS} \quad (2.4)$$

2.1.2. Haddock

Haddock is a data-driven docking method, in which either known biophysical information or theoretically predicted information about the protein-protein interface drives the docking process.[59,60]

The program requires of a list of residues directly implicated at the interface (active residues) or potentially implicated on it (passive residues) to start the process. Ambiguous Interaction Restraints (AIRs) are applied over the selected residues. An AIR can be defined as an ambiguous intermolecular distance, at a maximum of 3 Å, between each active residue of one protein and the rest of the selected residues (active and passive) of the partner protein. An effective distance between pairs of atoms determines an AIR energy term calculated through soft- square harmonic potentials:[57]

$$d_{iAB}^{eff} = (\sum_{m_{iA}=1}^{N_{atoms}} \sum_{k=1}^{N_{res}B} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}})^{(-\frac{1}{6})} \quad (2.5)$$

Where N_{atoms} is the number of atoms of a given residue, N_{res} indicates the sum of active and passive residues of a protein. A sum averaging of $1/r^6$ is used to emulate the attractive term of a Lennard-Jones potential, this guarantees that the restraints are satisfied as two atoms of two proteins are in contact.

The first step in the docking protocol consists of a rigid-body energy minimization.[58] In the first four cycles the proteins are allowed to rotate, then to rotate and translate and finally the pair stays docked. From there, the 200 best conformations in energy terms pass to a second stage of three cycles of simulated annealing refinement or semi-flexible refinement. Finally, in the third stage, the complexes are subjected to three short molecular dynamics runs. In the first, all the atoms are restrained except for those on the interface, in second the atoms at the interface are restrained, and in the third the backbone atoms outside the interface are restrained.[57-58]

In every stage, each docking conformation is scored and ranked based on average energy terms electrostatic, van der Waals and restraints violations and their average buried surface

area. The final structures are then clustered and analysed according to their average energy and then ranked.[57-58]

2.1.3. Protein-protein interface predictors

Knowing where on the surface protein-protein interaction occurs can limit the docking results and increase the chance of obtaining the correct conformation. Since the first protein-protein docking algorithm was published, many efforts have been made to theoretically predict the protein-protein interface.[52] Even though, experimental studies suggested that the binding site of Sap-C should be in the proximities of the residue N370,[42] as its mutations affects the association, in our research we have resorted to one interface predictor in order to dismiss other possible sites of interactions.[61]

CPORT

CPort is a consensus method for protein-protein binding site prediction, which combines the search of six different platforms for interface search.[62] It makes use of different free algorithms, which combines and optimises the results. It uses two structure-based algorithm: Promate[63] and SPPIDER[64], one conservation based algorithm: WHISCY, a neural network Cons-PPSI,[65] and one empirical Scoring Function based: PINUP[66]. This algorithm was designed in the same group as Haddock, and it has been extensively reported to work better than all the rest separately.[60]

2.2. Experimental

2.2.1. Protein-Protein Docking with Hex.

Docking Optimization

Two series of docking experiments were conducted to optimise parameters to perform protein-protein docking. In the first three experiments geometrical parameters were adjusted, which included angular and distance ranges. Those parameters that yielded better poses, in terms of distance between the partner proteins, number of electrostatic interactions and relative orientation of both proteins, were chosen to carry out a second series of optimization experiments, in which correlation type and post-processing procedure were evaluated in their different combinations.

In Hex, the larger protein is defined as the Receptor, while the smaller protein is the Ligand.[67] Here we have treated GCase as receptor and Sap-C as ligand. In order to limit the geometrical search of docking poses, Hex allows the user to adjust three geometrical parameters. Firstly, the distance range, which is the distance between the centroids of both proteins or the length of an intermolecular axis linking both centroids. Secondly, the angle range of the ligand and thirdly the angle range of the receptor that determines the degree of rotation of the protein relative to the defined axis. The range angles create imaginary cones that limit the rotational movement of the proteins.

Experiment	Receptor Range (°)	Ligand Range (°)	Distance Range (Å)
PARAMETER SET 1	45	45	20
PARAMETER SET 2	**60**	**60**	**40**
PARAMETER SET 3	75	75	40

Table 2.1: *First series of docking optimization experiments. Combinations of ranges that that produced best poses have been highlighted in red.*

As previously explained, Hex implements different algorithms for ranking solutions, which are termed as correlation types. Hex also offers multiple options of post-processing the docks. In the second series of experiments, different combinations of correlation types and post-processing parameters were used to optimise the results. A method unique to Hex employs the use of Decoys as a Reference State (DARS) to complete the docking information. DARS consists in a geometrical search after the interactions are compared and evaluated with decoys based on frequency of interactions.[68] This method can also be implemented after correlations have been made using other ranking algorithms.

Experiment	CORRELATION TYPE	POST-PROCESSING
PARAMETER SET 1	Shape	No post-processing
PARAMETERSET 2	**Shape**	**Bumps**
PARAMETER SET 3	Shape	OPLS E
PARAMETER SET 4	Shape	OPLS MM
PARAMETER SET 5	Shape	DARS E
PARAMETER SET 6	Shape	DARS MM
PARAMETER SET 7	**Shape + Electrostatics**	**Bumps**
PARAMETER SET 8	**Shape + Electrostatics**	**OPLS E**
PARAMETER SET 9	**Shape + Electrostatics**	**OPLS MM**
PARAMETER SET 10	Shape + DARS	Bumps
PARAMETER SET 11	Shape + DARS	DARS E
PARAMETER SET 12	Shape + DARS	DARS MM
PARAMETER SET 13	**Shape + Electrostatics + DARS**	**Bumps**
PARAMETER SET 14	**Shape + Electrostatics + DARS**	**OPLS E**
PARAMETER SET 15	**Shape + Electrostatics + DARS**	**OPLS MM**

Table 2.2: *Summary of the second docking calibration experiments. Parameter Sets selected and chosen to be further refined has been highlighted in red.*

These parameter sets resulted in docking poses where the relative location/orientation of the two proteins, and where a number of plausible side chain interactions were made, were selected to perform the rest of docking experiments.

Docking experiments were conducted in combinations of different conformations of both partner proteins. For example, apo GCase (PDB id 1OGS[25] and 2NSX.a[27]) and activated GCase (PDB id 2NSX.d[27]) were docked with closed (PDB id 2GTG[38]) and open conformation (PDB id: 2QYP[69]) of Sap-C. For each combination, two series of seven docking runs were carried out. In the first series, the centre of masses of each protein was used as a centroid or origin for all geometrical calculations. In the second series, residue H365 was selected as origin for GCase. The selection of H365 was based on the results of CPort algorithm and the results of experimental studies that locate Sap-C binding site in the proximities of the important residue N370. Using H365 as a centroid helped to narrow down the search in the GCase binding site, thus making the search more efficient (avoiding to test poses outside the binding site). It is important to note that the crystal structure with PDB id 2NSX has been used to test both Apo-GCase (2NSX.a) and activated GCase (2NSX.d). This is because the molecules in the asymmetric unit of this structure are different. The suffix at the end of the PDB id denotes the chain that has been used from the structure.

Experiment	Receptor Range (°)	Ligand Range (°)	Distance Range (Å)
PARAMETER SET 1	45	45	20
PARAMETER SET 2	**45**	**60**	**40**
PARAMETER SET 3	45	75	40

Table 2.3: *Second series of docking experiments, in which the centroid of the receptor protein was changed to residue H365. Calibration of geometrical parameters was recalculated for this series.*

Experiment	CORRELATION TYPE	POST-PROCESSING
PARAMETER SET 1	Shape	Bumps
PARAMETER SET 2	Shape + Electrostatics	Bumps
PARAMETER SET 3	Shape + Electrostatics	OPLS E
PARAMETER SET 4	Shape + Electrostatics	OPLS MM
PARAMETER SET 5	Shape + Electrostatics + DARS	No post-processing
PARAMETER SET 6	Shape + Electrostatics + DARS	Bumps
PARAMETER SET 7	Shape + Electrostatics + DARS	OPLS E

Table 2.4: *Parameters used for the first and second series of docking runs. A total of seven docking runs were conducted, each one with a different parameter set. In the first run the natural centres of masses of both molecules were conserved as centroid. In the second run the residue H365 was used as a centroid of GCase, whereas the natural centre of masses was used as a centroid of Sap-C.*

2.2.2. Protein-Protein Docking with Haddock

Six docking experiments were conducted using the "easy interface" of Haddock Server. Such interface allows the user to define the residues on the contact interface of both proteins. User may define "active residues" which will be taken to form a part of the interface and "passive residues" which will be considered to potentially be part of the interface.[60]

GCase	GCase- Active Residues	GCase- Passive Residues	SapC	SapC- Active Residues	SapC- Passive Residues
2NSX.d	H365	Auto	2GTG	V49, T52, Y53, S56, S57, I60	Auto
1OGS	H365	Q362, T369, Y373	2GTG	V49, T52,Y53, S56, S57, I60	K25
2NSX.d	H365, T369	Auto	2GTG	No	V49, T52, Y53, S56, S57, I60
1OGS	H365, T369	Auto	2GTG	No	N20, N21, K22, E24, E26, D29, S41, V49, T52, Y53, S56, S57, I60
2NSX.d	H365, T369	Auto	2QYP	No	N20, N21, K22, E24, E26, D29, S41, V49, T52, Y53, S56, S57, I60
1OGS	H365, T369	Auto	2QYP	No	N20, N21, K22, E24, E26, D29, S41, V49, T52, Y53, S56, S57, I60

Table 2.5: *Summary of the docking experiments carried out with Haddock. The active and passive residues have been chosen on the basis of the experimental data and predicted protein-protein interface from CPORT.*

2.2.3. Screening of docking orientations

The screening of docking poses was conducted in several steps and was applied to the best 20 solutions of each run:

1. The quality of the docks was assessed, by taking into account the relative position of one protein with respect to the other. For example, poses in which Sap-C was encroaching the active centre or was too far away from it were rejected. This was carried out through simple visual inspection.

2. Interaction of the docked Sap-C with constituent amino acids present in the helix 7 (of GCase) containing N370 was essential. Docks where this interaction was absent were rejected.

3. Number of interactions between the two proteins was analysed in detail. Poses with less number of interactions were rejected.

4. Only those docks in which Sap-C interacting residues corroborated with the experimental data were selected.

The poses that satisfied the above mentioned criteria were energy minimised.

2.2.4. Energy Minimisation with Amber.

Hex performs a rigid body docking. Hence all resulting docking poses presented steric clashes between both proteins side chains, even after using post-processing procedures. In order to overcome this problem, the docks were energy minimised using AMBER12[70] software.

The procedure used to minimise the structures was as follows:

1. Create molecular topology/parameter and coordinate files with LeAP, using FF99SB all atoms force field for complexes with explicit solvent and counter-ions.

2. Energy minimization of waters and counter ions by holding the protein complex fixed with positional restraints of 500 Kcal/mol, (1000 steps of minimisation).

3. Energy minimization of the complete system without any restraints (2500 steps of minimisation).

2.3. Results

2.3.1. Protein- Protein Interface Predictor

The program used for predicting the protein-protein interface identifies the binding site of Sap-C on GCase to be between Helices 6 and 7 of Domain III and Domain II. It is interesting to note that N370, is present on helix 7. This is consistent with experimental studies and is one of our criteria in selecting correct docked poses.

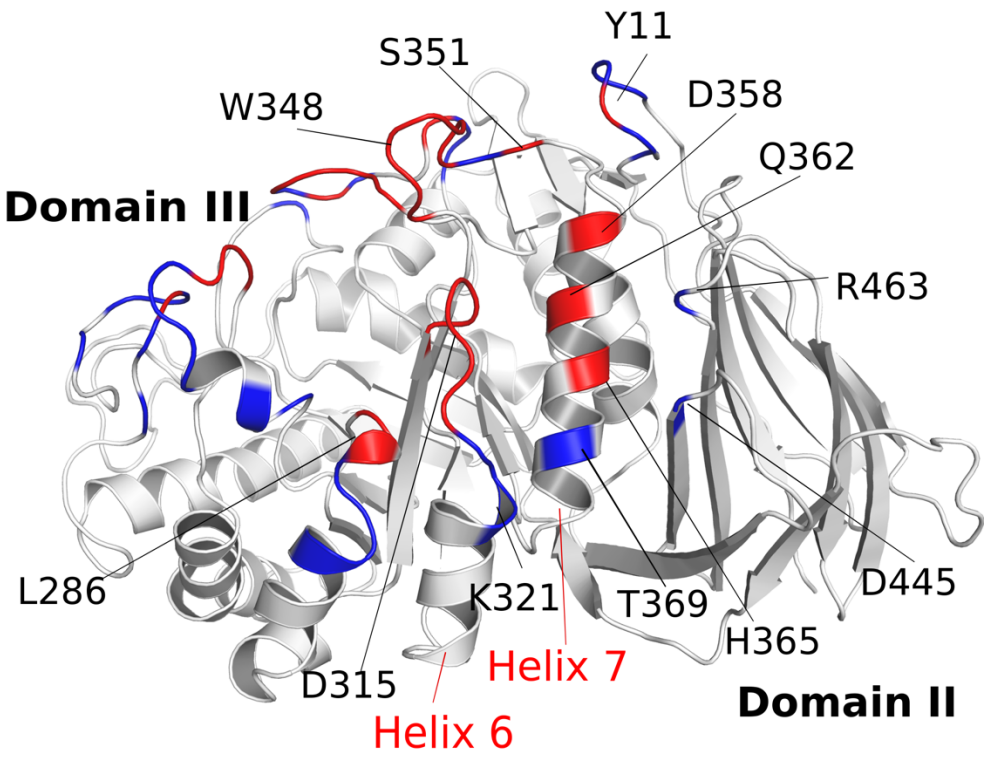

Figure 2.2: *The predicted protein-protein interface from CPORT algorithm. Residues in red are those considered by the program to take part in the protein-protein binding, and those marked in blue can potentially intervene in the bind. The program identified the protein-protein binding site over helix 7 of Domain III flanked by helix 6 and Domain II.*

2.3.2. Docking Results

3.3.2.1. GCase + Sap-C (Closed)

Hex

After applying the screening criteria to the top 20 clusters of output solutions for each docking, only six docked orientations of GCase and closed Sap-C were selected for energy minimisation.

As mentioned in the introduction, the main difference between Apo- (inactive) and active structure of GCase is the different conformation of Loop-1. In inactive Apo-GCase, Loop-1 shows an extended conformation whereas in active GCase the loop is in helical conformation. Except for Loop-1, the differences between apo- and active GCase are minimum. The cavity predicted to be the protein-protein binding site is similar in Apo- and active GCase. (The Cα-RMSD of the alignment between Apo-GCase (1OGS) and active GCase (2NSX.d) is 0.353 Å. The Cα-RMSD of the alignment between Apo-GCase (2NSX.a) and active GCase (2NSX.d) is 0.348 Å).

The docking poses selected for energy minimization were aligned one against the other. It was seen that some poses could be considered similar. Although obtained with different combination of proteins (namely Apo-GCase and closed Sap-C and active GCase and closed Sap-C) the relative position of both proteins was almost identical, with the RMSD of the alignment of both poses being less than 0.6 Å. Table 6 summarises the poses considered for energy minimisation.

	MODEL	Series	Parameter set	Docking Pose (cluster number)
Pose 1	2NSXa-2GTG	1	5	4
Pose 2	1OGS-2GTG	2	2	8
Pose 3	1OGS-2GTG	2	2	1
Pose 4	2NSXd-2GTG	1	6	3
Pose 5	2NSXd-2GTG	2	1	1
Pose 6	2NSXd-2GTG	2	1	8

Table 2.6: *Summary of the six poses (models) selected for energy minimization. Pose 1 and 3 and Pose 2 and 5 were almost identical.*

MODEL	GCase- Residues	Sap-C Residues	HB and Ionic pairs
2NSXa-2GTG-pose 1 1OGS- 2GTG-pose 3	F316-P319, K321, Q362, HY365, T369, Y373, K441, D443, R463-S465	V14, L17, N21-T23, E26, I27, L29, A30, K33, M34, K37, E48	K33-T369, D30-H365, K26-Q362, E14-K441
1OGS-2GTG-pose 2 2NSXd-2GTG-pose 5	F316-P319, K321, K346, W348, R353, D358, Q362, H365, T369, Y373,D443, K441, R463-S465, Y487	K25, E26, L29, A30, K33, M34, K37, S41, E44, E45, E48, T52, Y53, I57, V60	K26-T369, D30-H365, K33-D315, Q48-D358, E25-K441
2NSXd-2GTG-pose 4	F316-P319, Y321, K346, W348, R353, S356, D358, Q362, H365, T369, K441, R463, Y487	C4, E26, L29, A30, K33, K37, S41, E45, C46, E48, V49, T52-G54, L76	T53-H365, E49-R463, E45-S356
2NSXd-2GTG-pose 6	F316-P319, Y321, K346, W348, R353, D358, Q362, H365, T369, Y373, K441, D443, Y487, R463-S465	T23, I27, 2E6, A30, F31, M34, K37, S41, E48, T52, S56	K26-Y373, D30-Y369, D33-H365

Table 2.7: *Summary of the interacting residues in the selected models and relevant electrostatic interactions for selected poses of GCase and closed conformation of Sap-C.*

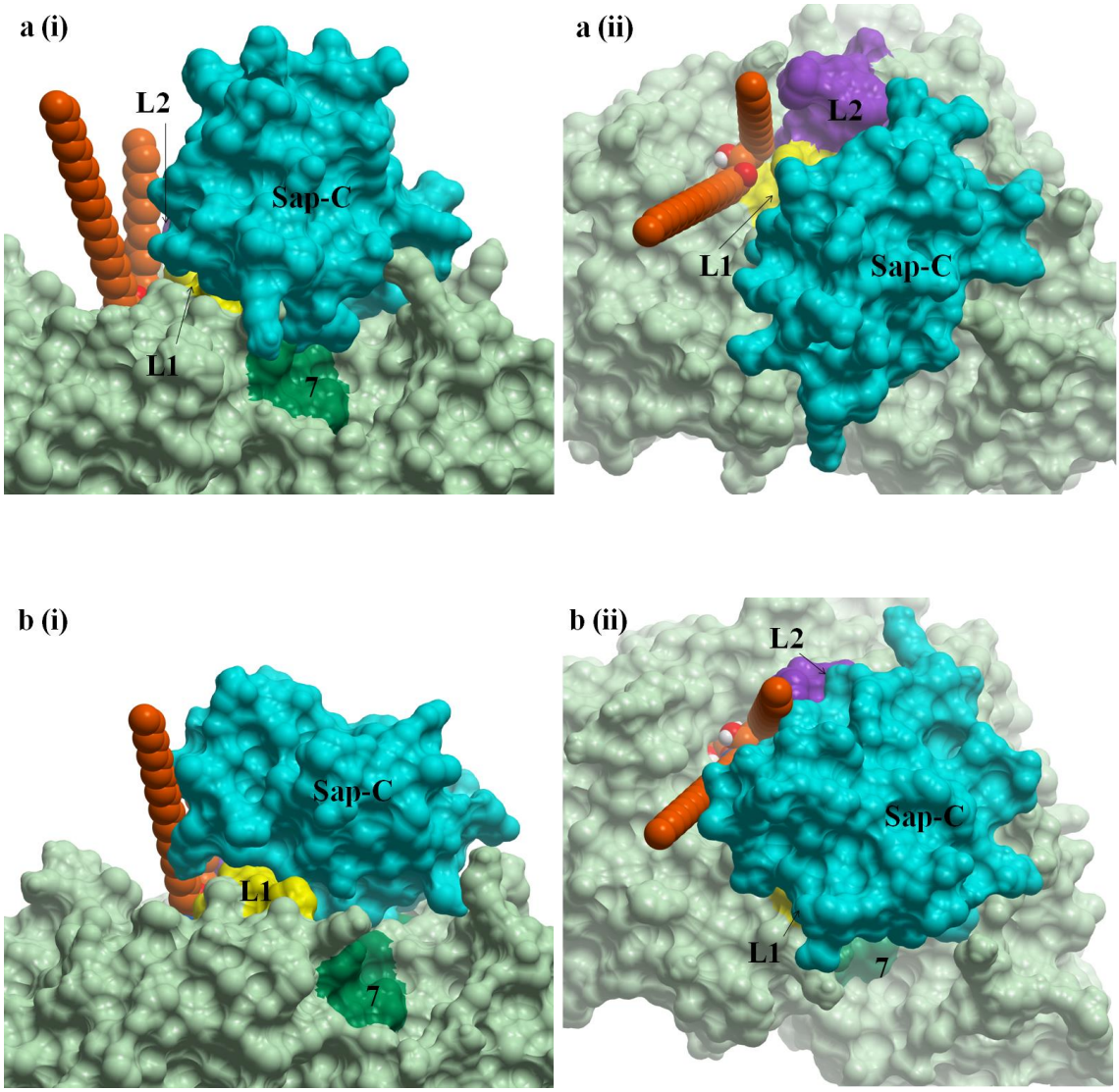

Figure 2.3 A: *Surface representation of the top docking poses that fulfil the selection criteria. Each docking pose has been depicted in two perspectives. Docking poses 1 and 2 have not been included as they are same poses as 3 and 5. GCase is shown in light green. Loop-1 at the entrance of the active site of GCase has been coloured in yellow and Loop-2 in purple. Helix 7 of GCase has been coloured in green. Sap-C is shown in cyan. GluCer has been represented as orange spheres. Pose 3 has been illustrated in* ***(a)****, Pose 4 has been illustrated in* ***(b)***

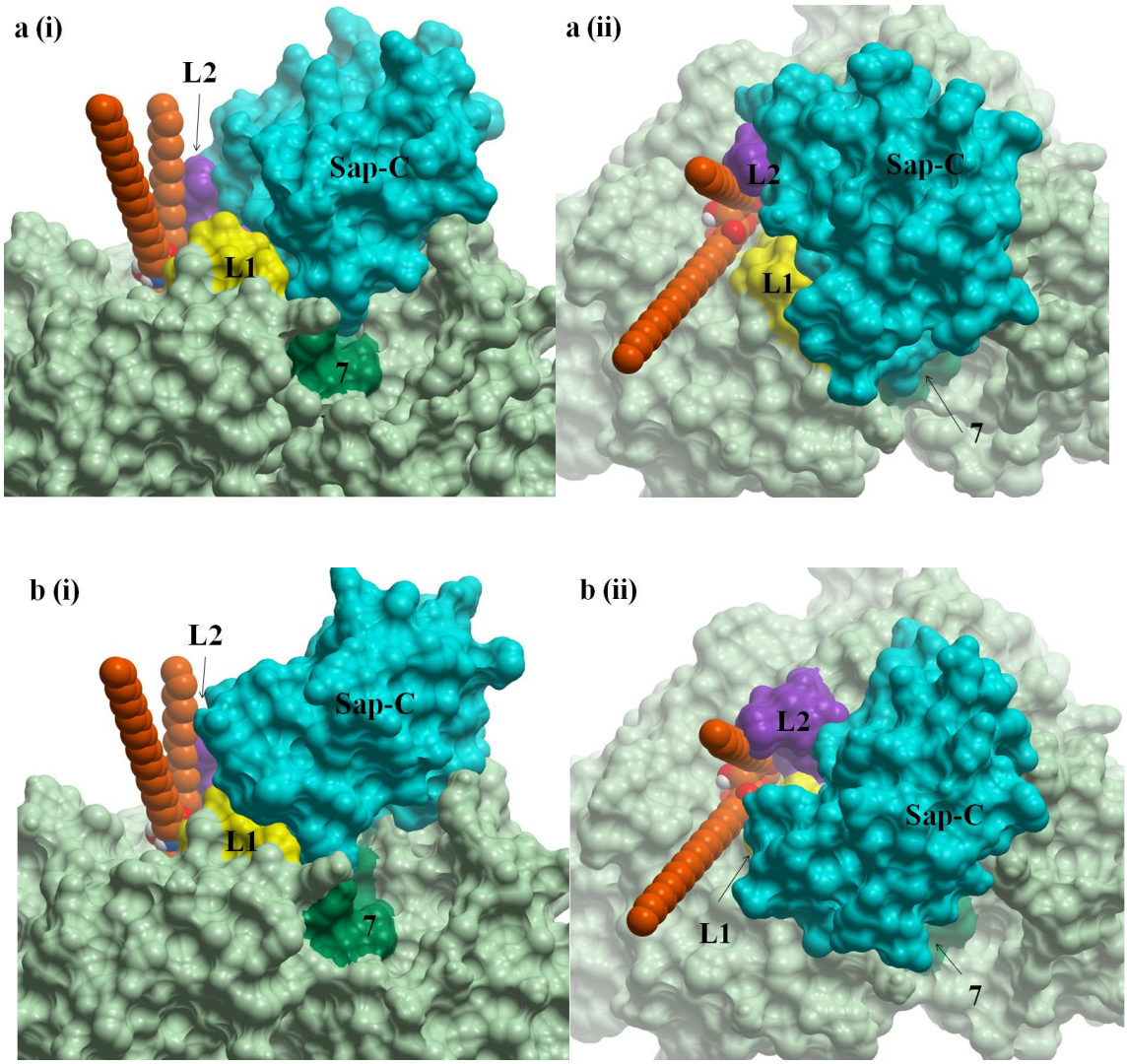

Figure 2.3 B: *Surface representation of the top docking poses that fulfil the selection criteria. GCase is shown in light green. Loop-1 at the entrance of the active site of GCase has been coloured in yellow and Loop-2 in purple. Helix 7 of GCase has been coloured in green. Sap-C is shown in cyan. GluCer has been represented as orange spheres. Pose 5 has been illustrated in* ***(a)*** *and Pose 6 has been illustrated in* ***(b).*** *In the best pose, Pose 5, Sap-C fills the cavity formed by Loop-1 and 2, helix 7 and Domain II of Sap-C.*

Haddock

None of the poses obtained from the first two runs of Haddock passed the first two selection criteria. Two poses from the third run passed the first two criteria but were not in agreement with the experimental data as both N- and C-terminal of Sap-C formed part of the binding site. It was when the predicted active and binding domains of Sap-C (N20, N21, K22, E24, E26, D29, S41, V49, T52, Y53, S56, S57, I60) were coerced as passive residues when we observed one of the poses already obtained with HEX. The identified dock was very similar to the docking poses 2 and 5 from HEX experiments. The RMSD of the alignment was 0.62Å.

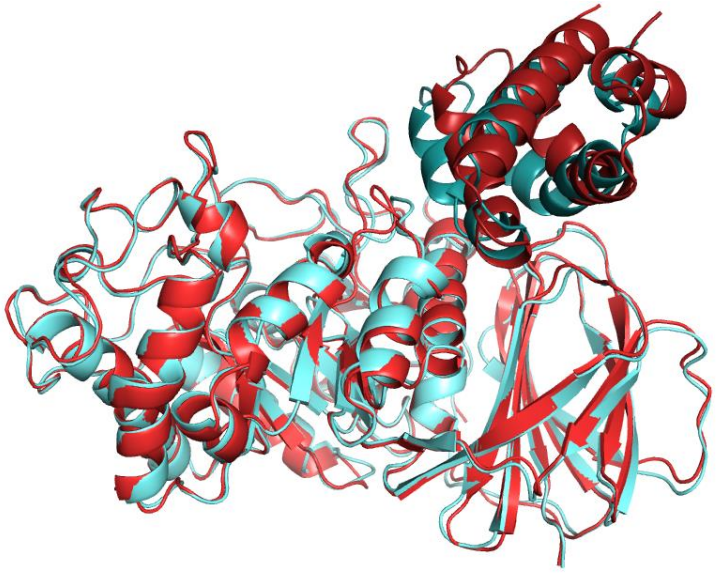

Figure 2.4: *Superposition of the complexes obtained from Hex- pose 2 (second series, eighth docking pose) in light blue and one of the poses obtained from the third run of Haddock in red. Both complexes share the same relative position of the pair of proteins.*

Due to the consistency with experimental data, the correspondence across different docking algorithms and the concordance of the binding mode when two different conformations of GCase were used, we selected Pose-2 (extended) and Pose-5 (helical), for further studies.

2.3.2.2. GCase + Sap-C (Open)

Hex

We first analysed the results of the docking experiments that were run using the open conformation of Sap-C. Even though the final pose that was selected was the only pose that satisfied experimental data, as well as the only one found using two different docking methods and two different conformations of GCase, we were receptive to find new and better models. We also identified an equivalent pose with the open conformation of Sap-C, in some of the docking runs. Finally, we selected two docking poses to represent our model, which are summarised in Table 2.8 and 2.9.

	MODEL	**Series**	**Parameter set**	**Docking Pose (cluster number)**
Pose 7	1OGS- 2QYP	1	2	3
Pose 8	2NSX.d-2QYP	1	5	7

Table 2.8: *Summary of the two poses (models) selected for energy minimization. Pose 1 and 2 were almost identical.*

MODEL	**GCase- Residues**	**Sap-C Residues**	**HB and Ionic pairs**
Pose 7- 1OGS- 2QYP Pose 8- 2NSX.d-2QYP	D315- K321, K346-E349, R353, D358, Q362, H365, Y373, D443, K441, D443- D445, R463-S466, Y487	K25, E26, L29, A30, K33, M34, K37, S41, E44, E45, E48, T52, Y53, I57, V60	Q48-W348, K26-Q362, D30- H365, E25-K441, D443- S60, N22- D445, S57- 466

Table 2.9: Summary of the interacting residues in the selected models and relevant electrostatic interactions for selected poses of GCase and open conformation of Sap-C.

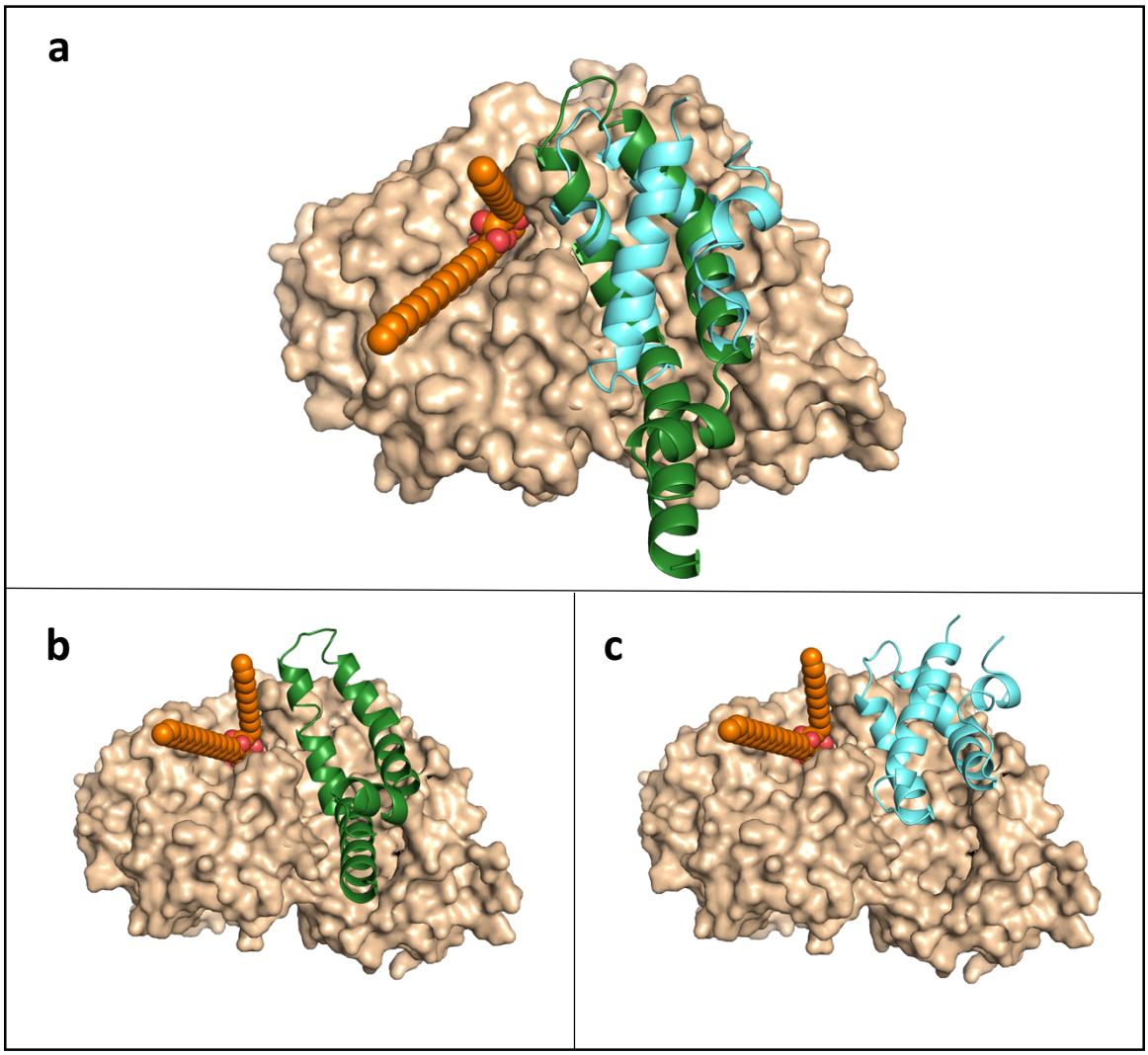

Figure 2.5: *(a) Alignment of 2NSX.d-2GTG-pose 5 and 2NSX.d-2QYP-pose 8. The figures illustrates GCase in complex with Sap-C in open (green) and closed (cyan) conformation. GCase has been depicted in surface representation and coloured in light brown. (b) 2NSX.d-2QYP-pose 8, GCase (light brown) in complex with closed Sap-C (green). (c) 2NSX.d-2QYP-pose 8, GCase (light brown) in complex with closed Sap-C (cyan). GlyCer is drawn as orange spheres.*

2.4. Discussion

The GCase-Sap-C model presented by Atrian et al. in 2008 is valuable, although not in complete agreement with the experimental data.[46] Besides, this model is not available in the public domain. In the work presented in this thesis, we have followed a knowledge-based docking protocol to identify the GCase-Sap-C protein-protein interface.

It is important to point out contradicting data between experimental studies, on Sap-C interface with GCase. Weiler et al. based their study on competition of synthetic peptides derived from Sap-C, located two binding sites at position 6-27 and 45-60, and two activation sites at position 27-34 and 41-48 on Sap-C[45], judging by the capacity of these peptides to bind or activate GCase hydrolysis of a fluorescence GCer analog (4-methylumbelliferyl-β-D-glucoside. Another study by Qi et al., based on chimeric saposins, pointed just one activation site between residues 47- 62.[43] Spatially positioning the sites on the structure of Sap-C, we immediately observe that all the data cannot be reliable. Binding site 1 and 2 from Weiler et al. is not compatible as they are on opposed edges of the protein. Comparing the results of the two studies we inferred that:

1. There is an observed coincidence in the predicted activation site proposed by Qi et al. and the second binding site observed by Weiler et al. In fact, they are same. So if we have to choose one binding site then it has to be the preferred one.
2. The second binding site identified by Weiler et al. binds with stronger affinity, so is more likely to be the binding site.
3. The first binding site is not compatible with the membrane anchoring experiments (see chapter 3).

Based on these criteria, we started looking for the second binding site suggested by Weiler et al (residues 45-62) on helix 3 of Sap-C. It was most likely to be the binding interface with GCase. Furthermore, the two activation sites proposed by Weiler (helix 2 and 3) and that proposed by Qi (helix 3 and 4) are topologically compatible with the second binding site. We therefore conclude that helices 2, 3 and 4 of Sap-C should constitute the protein-

protein binding interface.

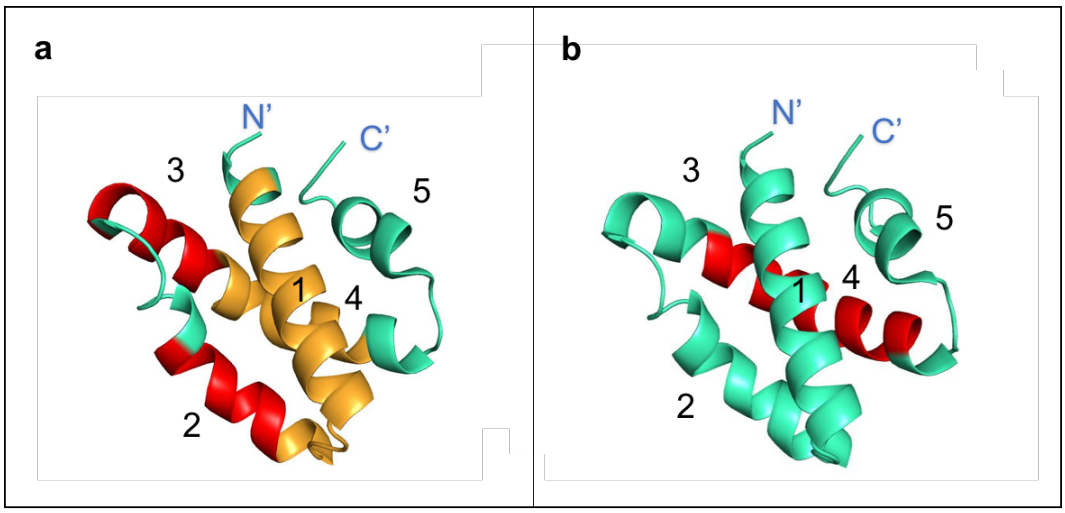

Figure 2.6: *The structure of Sap-C.* ***(a)*** *The binding site (residues 6-27 and 41-60; orange) and the activation sites (27-34 and 40-49; red) as proposed by Weiler et al. and* ***(b)*** *The activation site as proposed by Qi et al. (47-62; red) has been illustrated. Please note that the protein Sap-C is composed of 5 helices and not just four as it could seem at first glance. Helices 3 and 4 are separated by a kink instead of a loop.*

One can now think of two possibilities for selection based on the experimental observation of activation by Sap-C:

1. The two activation sites identified by Weiler et al. lie adjacent to the loops at the entrance of the active site of GCase and exerts its actions on the surrounding environment.

2. The activation site of Sap-C defined by the Qi et al. lies adjacent to the loop at the entrance of the active site of GCase.

At first instance, we thought that the second option would probably be the answer. We thought so because the activation site proposed by Qi et al. coincided with the second binding site proposed by Weiler et al. We oriented the proteins such that the activation site proposed by Qi et al. was positioned in the proximities of the loops at the entrance of the active site. We also defined this site as active and passive residues in the Haddock program

(see Table 2.5). We identified only one pose that was consistent with our selection criteria. However, we were unable to identify a similar docking pose via Hex.

A pose in agreement with the first criteria was identified via Hex. It was found in both cases, when activated and apo- conformations of GCases were used. This pose was also identified in results from Haddock, albeit being the only plausible pose. Interestingly, the pose was obtained with Haddock only after we stopped coercing a binding site that satisfied the second criteria (see Table 2.5). We also found a GCase-(Open) Sap-C model in this orientation.

Apart from the aforementioned experimental data about the binding and the activation site in Sap-C, we relied on the solved crystal structure of a Galactocerebrosidase (Galc) and Saposin A (Sap-A) complex.[71] Galc is another lysosomal hydrolase that catalyses the cleavage of galactocerebrosides. The overall fold of Galc is very similar to the structure of GCase, and it comprises of three domains: Domain I, a $(\alpha/\beta)_8$ TIM Barrel containing the binding site equivalent and highly similar to Domain III in GCase; Domain II, a β-sandwich that present a similar topology to Domain II in GCase; and Domain III, a lectin Domain that has no equivalence in GCase.

The protein binding site in Galc lies between Domain I and III, within the proximities of the active site.[69] Although Galc shares an overall common structure with GCase there are some key features that make the protein surfaces rather different. Firstly, one of the loop of the β – sandwich domain wrap over the active site constituting, probably, an activation loop. This loop changes the equivalent cavity formed in GCase between helices 6 and 7 and Domain II. Secondly, Domain III (lectin domain), absent in GCase, provides support for the binding of the cofactor, drawing a subtle cavity with some of the loops of the N-terminal side of the TIM Barrel. GCase lacks this cavity. Thirdly, even though both enzymes share the TIM Barrel motif, the loops at the entrance of the binding site are rather different in both proteins. The longest loops in Galc are found in the C-terminus side of the molecule, whereas the catalytic residues (E258 and E182) are exposed.[72] Figure 2.8 illustrates the crystal structure of the complex alone and aligned with our model of GCase

and Sap-C.

In 2006, the crystal structures of Sap-C and Sap-A were solved in a study carried out Victoria E. Ahn et al.[38] This study pointed out the differences in the electrostatic surface of both proteins, attributable to their different function. In 2001, Gregory A. Grabowski had already reported the different lipid binding properties of both proteins and correlated them with a different mechanism of action and different activation domains.[73] The different binding modes to the membrane are shown in Figure 2.7.

The differences in the structure and electrostatic surface between GCase and Galc, and Sap-C and Sap-A, the disparity in the activation domains of both cofactors, as well as the different binding properties, made us think both complexes, namely, Galc-SapA and GCase-SapC could be somehow different although with some common features. Those shared characteristic of the protein-protein binding site would be for example, lying beside the active site of the enzyme without encroaching it, being in agreement with the membrane interaction experiments or with the experimental determination of the activation domain.

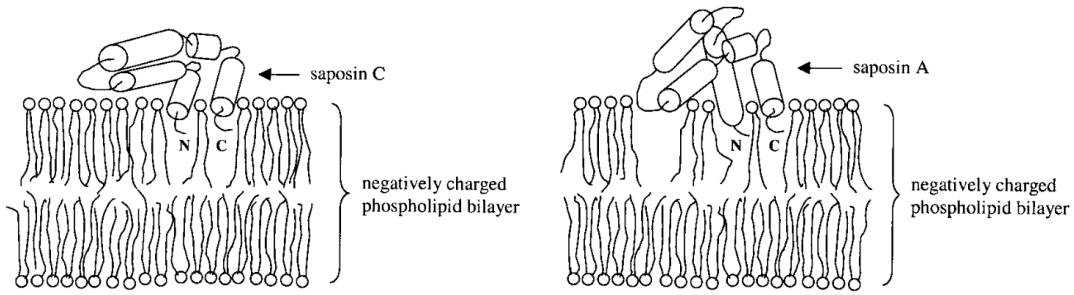

Figure 2.7: *The membrane binding mode of (a) Sap-C and (b) Sap-A. It can be seen that, even though both proteins have a high structure similarity, their mode of interaction with the membrane is different. Taken from reference 71.*

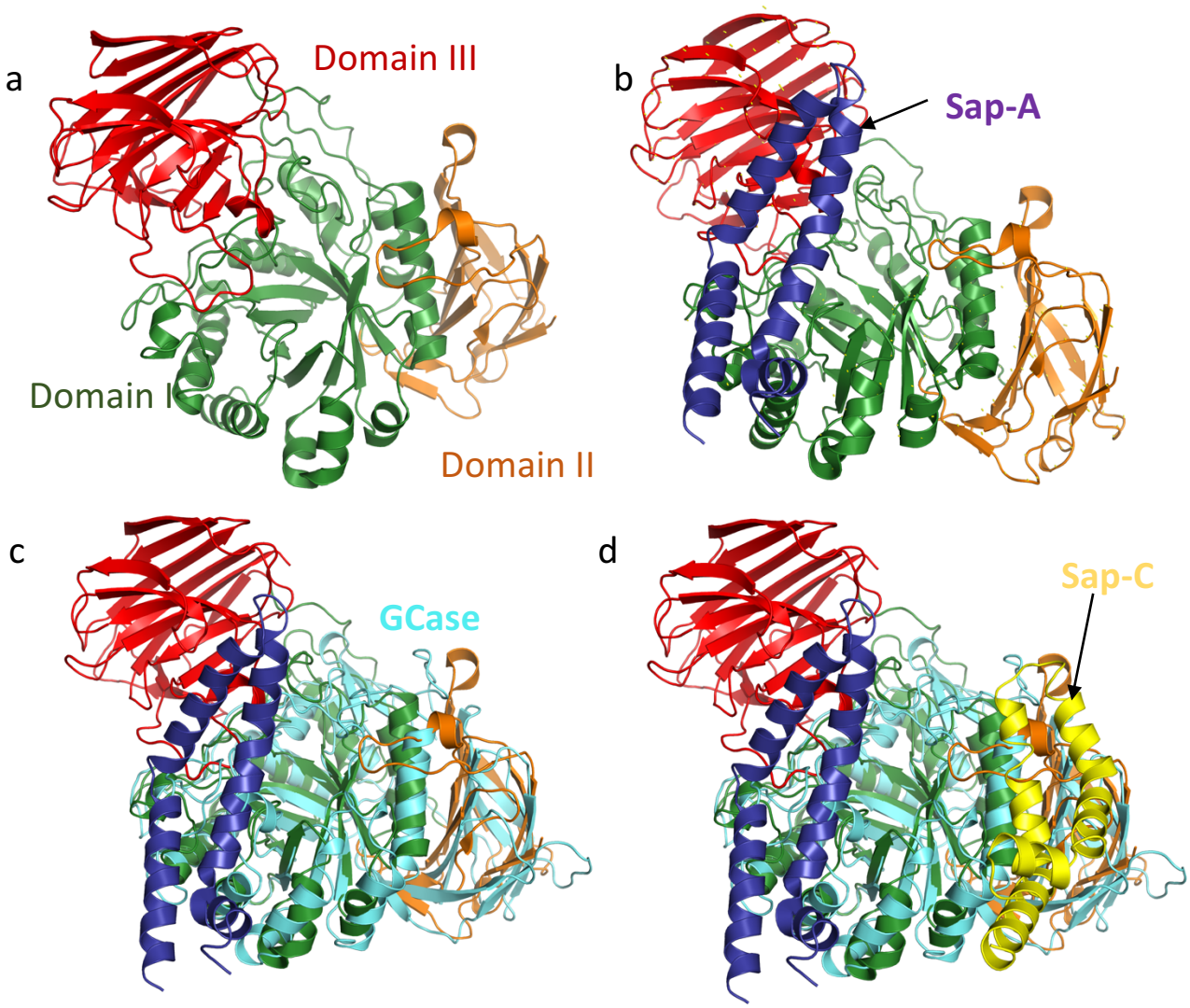

Figure 2.8: (*a) Crystal structure of Galc (PDB ID: 5NXB*[71]*). Domain I is depicted in green, Domain II in orange and Domain III in red. (b) Crystal structure of Galc in complex with Sap-A (purple) (PDB ID: 5NXB). (c) The crystal structure of Galc in complex with Sap-A has been aligned with the crystal structure of GCase (cyan). (d) The crystal structure of Galc in complex with Sap-A has been aligned with the crystal structure of GCase (cyan) in complex with Sap-C (yellow).*

Here we present this protein-protein model that has been obtained using the first criteria (based on the two activation sites identified by Weiler et al., lie adjacent to the loops at the entrance of the active of GCase and exerts it actions on the surrounding environment) and that in our opinion fulfils all the requirements to be a good pose and it is in agreement with the experimental data.

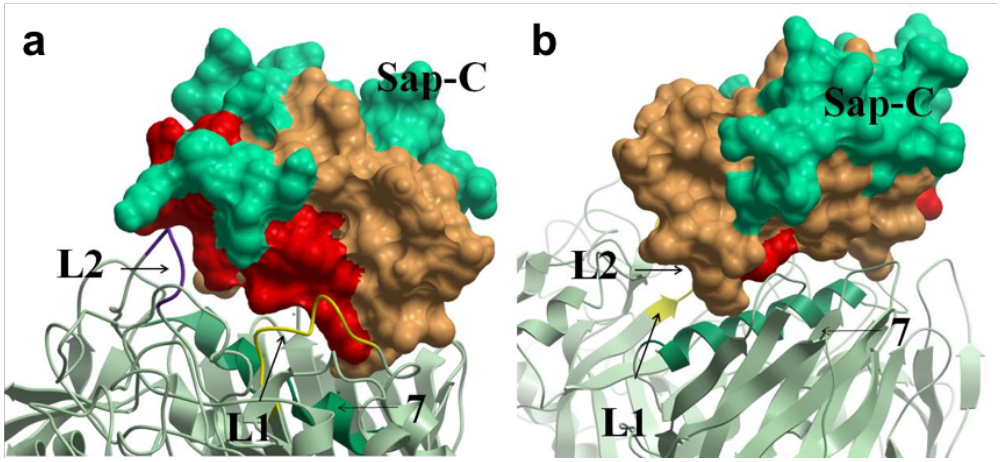

Figure 2.9: *(a-b) Two different views of the predicted GCase-Sap-C complex. GCase is shown as a cartoon in light green: Loop-1 in yellow, Loop-2 in purple and Helix 7 in dark green. Sap-C is shown as a surface, the binding site (residues 6-27 and 41- 60; orange) and the activation sites (27-34 and 40-49; red) as proposed by Weiler et al. and the rest of Sap-C in turquoise.*

CHAPTER 3:

MOLECULAR DYNAMICS

CHAPTER 3: MOLECULAR DYNAMICS

3.1. Introduction

The Function of biological macromolecules stems from their three-dimensional structure. Molecular dynamics (MD) simulations provide a valuable way to understand the physical basis of conformational changes that macromolecular structures undergo and that, ultimately, lead to the their biological function.[74,75] The internal motions determined by the structure produce the conformational dynamics of proteins critical to their function. Therefore, the connection between spatial structure and dynamics is necessary for the comprehension of protein functionality.[72-76]

The traditional way to explore conformational changes in proteins, so as to understand their biological function, was to accumulate many experimental structures that would cover the conformational space.[77] The rapid development of computer power in the seventies, allowed to approach the proteins conformational changes in a dynamic way by letting all the atoms in the macromolecule interact for a period of time. Thus, molecular dynamics simulations started to provide a connection between structure and dynamics turning into an interesting tool to explore the conformational energy landscape accessible to biomolecules.[72,73]

Molecular dynamics simulations offer atomic detail of the evolution of a system as a function of time, difficult to reach through experimental techniques; thus this theoretical technique can provide a complete picture of the dynamic features of the system and answer specific questions about the structural mechanism of action of proteins.[78]

In this context, we used molecular dynamics simulations to understand the structural mechanism of function of the enzyme GCase, to explore its structural dynamics and conformational changes that promote or are promoted by the interaction with other components including its facilitator protein, lipid membrane, etc. The evaluation of the protein dynamics offered complete understanding of the role of each part of the protein and thus provided structural implications of different mutations. A molecular dynamics simulation is the only method that allows us to study the dynamic interactions between the

proteins GCase and Sap-C and them with the membrane.

3.1.1. Molecular dynamics: Theory

Molecular dynamics simulations could be used to study the evolution of the internal motions of a protein and can also produce detailed thermodynamic and kinetic information.[76] Molecular dynamics simulations provide information at the atomic level like velocities and coordinates of atoms, this information can be converted, via statistical mechanics, to macroscopic observables, quantifiable properties such as heat capacities, energy and pressure. Statistical mechanics provides the mathematical foundation to relate atomic motion and distribution data to macroscopic properties, usually defined in terms of time-independent statistical averages. Thus, statistical mechanics is able to predict macroscopic phenomena of a biological system from individual molecular properties.[72-74,79]

Since the development, in the twentieth century, of the theory of the quantum mechanics (QM), the dynamics of the particles can be described by a new equation of motion, namely, the Schrödinger equation.[77] Molecules could potentially be described in terms of interactions of the nuclei with the electrons, whereas molecular geometry can be described as energy arrangements of nuclei. The Schrödinger time- independent Equation could be solved for a hydrogen atom through following equation.[77]

$$\left[-\frac{\hbar^2}{2m}\nabla^2 - \frac{Z}{r}\right]\psi(R) = E\psi(R) \qquad (3.1)$$

In this equation, the term in the square brackets represents the potential and kinetic energy of the electron; the distance of the electron from the nucleus is represented by r and the charge of the nucleus by Z. ∇ is the Laplacian operator (sum of second derivatives of the function with respect to each independent variable). $\hbar$ is the reduced Planck constant ($\hbar = h/2\pi$). The ψ is the state or wave function representing the coordinates of the electron, the E represents the electronic energy in atomic units and the R represents electron coordinates. The wave function for hydrogen atoms are essentially the atomic orbitals: s, d, p …

The Schrödinger equation can be generalised to a multinuclear, multi-electron system, as follows:[77]

$$\mathrm{H}\Psi = \mathrm{E}\Psi \tag{3.2}$$

In above equation, the Ψ is the multi-electron wave function H is the Hamiltonian operator, which is:

$$\mathrm{H} = -\frac{1}{2}\sum_{i}^{electrons}\nabla_i^2 - \frac{1}{2}\sum_{A}^{nuclei}\frac{1}{M_A}\nabla_A^2 - \sum_{i}^{electrons}\sum_{A}^{nuclei}\frac{Z_A}{r_{iA}} + \sum_{i<j}^{electrons}\sum\frac{1}{r_{ij}} + \sum_{A<B}^{nuclei}\sum\frac{Z_A Z_B}{R_{AB}} \tag{3.3}$$

Where Z is nuclear charge, M_A is the ratio of mass of nucleus A to the mass of an electron, R_{AB} is the distance between the nuclei A and B, r_{ij} is the distance between the electrons i and j and r_{iA} is the distance between electron i and nucleus A.

However, the Schrödinger equation has never been solved for a multi-electron system, not even for a two electron system such as the hydrogen molecule or the helium atom.[75,77] So that, in order to describe the dynamics of a complex biological system, an approximation to the Schrödinger equation needed to be introduced. One of the most used methods to study the dynamics of biomolecules is the Born-Oppenheimer approximation.[80] This approximation to the Schrödinger equation relies on the dramatic difference of mass between electrons and nuclei.[81] The electrons are much lighter and hence move much faster than the nuclei, so that they rapidly adapt to the position of the nuclei. Thus, the motion of the electrons can be neglected allowing the dynamic to be decoupled in two subsystems, one slow subsystem, the nuclear, and one fast that follows the state of the former (electron). In practise, the use of this approach allows the nuclei and electrons to be quantified into a sole atom-like particle.[75-82]

3.1.1.1. Molecular Mechanics: Force fields.

The forces acting on atom-like particles are computed using Molecular Mechanics force fields. Unlike Quantum Mechanics, Molecular Mechanics quantify the energy of the particles based on their atomic entities, nuclei and electrons being unified.[83] Force fields are mathematical concepts that combine first-principles physics and parameter fitting to quantum mechanical calculations and empirical data to define molecules, represented as atoms connected by bonds with lengths, angles and energies.[84] There are different types of force fields developed with different levels of complexity designed to be used for different systems, schematically a force field being:

$$E_{TOTAL} = E_{Stretch} + E_{Angle} + E_{Torsion} + E_{VdW} + E_{Electrostatics} \quad (3.4)$$

This simplified equation represents the total sum of energies exerted over a particle in a macromolecule. The total energy is the sum of bonded ($E_{Stretch} + E_{Angle} + E_{Torsion}$) and non-bonded interactions ($E_{VdW} + E_{Electrostatics}$).[85,86]

Molecular Mechanics considers molecules as weights connected by strings.[86] The potential energy of the stretch and angles of the bonds are calculated through the Morse potential and Harmonic potential, respectively, the latter being derived from the Hooke's law (Harmonic Oscillator Model).[87]However, Morse potential is not normally used in molecular mechanics force fields, a simpler approach is to use Hook's law in which the energy changes are based on square dislocation from the reference bond length, which is known as l_0 and k being bond constant, as shown in equation 3.5.[83-85]

$$\upsilon(l) = \frac{k}{2}(l - l_0)^2 \quad (3.5)$$

The angle bending is treated the same manner as bond stretching using Hook's law through the deviation of the angles from a reference angle via the following equation: [83-85]

$$\upsilon(\Theta) = \frac{k}{2}(\Theta - \Theta_0)^2 \quad (3.6)$$

Where Θ_0 is the reference angle and k is the force constant. It is estimated that a smaller energy is required to move the angle from the reference or equilibrium than to stretch a bond. [83-85]

In bond stretching and angle bending the creation of any change or deformation from the equilibrium requires a large energy input, so that the main difference in structures and energies are due to the torsional and non-bonded interactions. The structural elements as well as the molecular geometry could be easily understood by defining the existence of barriers to rotation about bonds. The energy of torsion considers the periodicity of rotation in bonds. [83-85]

$$\upsilon(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] \qquad (3.7)$$

Torsional potentials are mostly represented as a cosine series expansions. V_n is the barrier height, γ determines the barrier where the torsion angle crosses its minimum value and the ω is the torsion angle. [83-85]

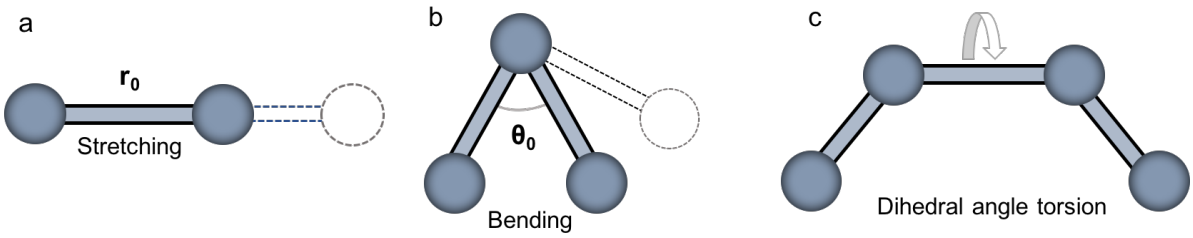

Figure 3.1: *represents the bonded interactions in the model of molecules as weights and bonds: (a) bond stretching, (b) angle bending and (c) dihedral angle torsion.*

The corresponding energies of van der Waals and electrostatic interactions are calculated by Lennard Jones Potential and Coulomb's Law respectively.[87]

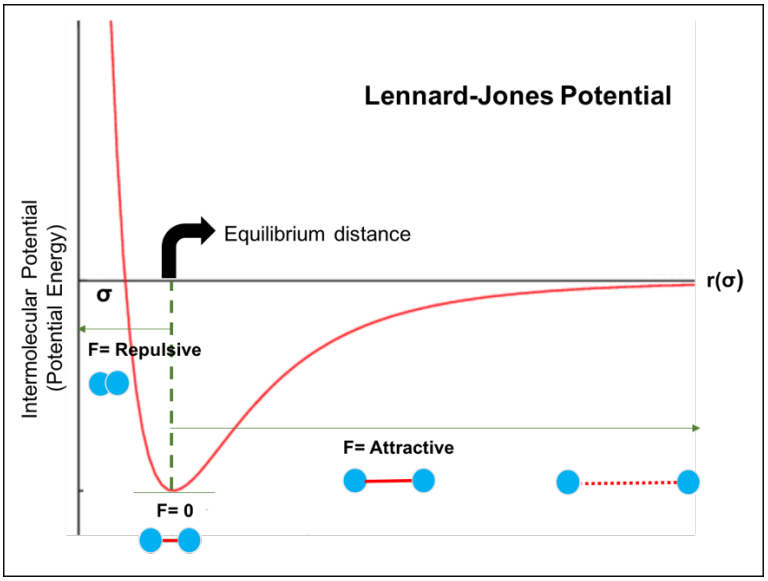

***Figure 3.2**: Potential Energy curve for the Lennard-Jones Potential, which represents the long range of attractive forces that holds atoms together.*

As seen in Figure 3.2, at the equilibrium distance the net force is equal to zero and the potential energy is minimum. For separation smaller than equilibrium distance repulsive forces start to take effect and the potential energy increases up to the collision diameter (σ).[88,89]

The electrostatic interactions represent a long range of attractive (or repulsive, depending on the charges) forces which become slowly weaker as a function of 1/r (Eq. 3.8). The Lennard Jones potential includes two terms one for long-range attractive interactions (van der Waals and dispersion interactions) that are function of $1/r^6$ and the other for the short-range of repulsive interactions $1/r^{12}$ (overlapping electron orbitals) (Eq. 3.9).[83, 85-87]

$$E_{ELECTROSTATIC} = \sum_{i}^{Natoms} \sum_{j>i}^{Natoms} \left(\frac{q_i q_j}{D r_{ij}} \right) \tag{3.8}$$

Equation 3.8 expresses Coulomb's law in which q_i and q_j are the atomic charges of the atoms i and j respectively, r is the distance between atoms and D is proportional to the dielectric constant of the material that surrounds the charges (e.g. water).

$$E_{VDW} = \sum_i^{Natoms} \sum_{j \neq i}^{Natoms} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \quad (3.9)$$

Equation 3.9 expresses the Lennard Jones potential, the parameters A_{ij} and B_{ij} are the van der Waals constants (determined experimentally or through modelling) and r_{ij} corresponds to the separation radius between the atoms i and j.[86,87]

So that, a typical potential energy function from a Molecular Mechanics force field would have the following form:[85]

$$U = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_\alpha (\theta - \theta_0)^2 + \sum_{torsions} \frac{v_n}{2} [1 + \cos(n\omega - \gamma)] + \\ + \sum_i^{Natoms} \sum_{j \neq i}^{Natoms} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{elec} \frac{q_i q_j}{D r_{ij}} \quad (3.10)$$

The equation 3.10 defines a general Potential Energy function as the implemented by a MD algorithm. Each one of the terms in this equation has been already explained thoroughly.

3.1.1.2. Molecular Dynamics Algorithm.

Molecular Dynamics solves Newton equations of motion to calculate the atom trajectories.[76] First, the potential energy is calculated for each atom. Then, the force to which each atom is submitted to, thus, obtained as a first derivative of the potential energy function with respect to the atomic position. The change in position of the atoms as a result of the forces exerted over them is calculated for each increment of time. The position is, therefore, obtained in each increment of time, atom trajectory being drawn along the simulation.[5,72-76]

Molecular dynamics algorithms take initial atom positions and initial distributions of velocities as input.[73-76] The force exerted over an atom can be calculated deriving the potential energy function with respect to the atom position:

$$F_i = -\frac{\partial U(R)}{\partial r_i} \quad (3.11)$$

As the second Newton's equation:

$$F_i = m_i\, a_i \text{ or } F_i = m_i\, \frac{\partial^2 ri}{\partial t^2}; \quad \text{Then, } -\frac{\partial U(R)}{\partial ri} = m_i\, \frac{\partial^2 ri}{\partial t^2} \tag{3.12}$$

U represents the potential energy based on the coordinates of the n atoms and the equation has to be solved numerically using a suitable algorithm.[74,5] This numeric solving is done through discretization of the trajectory and uses an integrator to move forward the trajectory over small steps or frames:

$$r_i(t_0) \rightarrow r_i(t_0 + \Delta t) \rightarrow r_i(t_0 + 2\Delta t) \rightarrow \cdots\ r_i(t_0 + n\Delta t) \tag{3.13}$$

The atomic position can be obtained for time step (Δt) using the following equation

$$r_i\,(t + \Delta t) = r_i\,(t) + v_i\,\Delta t + \frac{1}{2} a_i\,(\Delta t)^2 \tag{3.14}$$

Thus, atoms position is obtained every time step. All atoms positions in the protein are calculated simultaneously, for every time step. The variation in atomic positions along the time draws a trajectory, which ultimately allows us to know the evolution of the biological system along time. [5,72]

These integrators possess properties such as high accuracy, stability if large Δt is used and speed for force calculations. These algorithms are simple, efficient, stable and time-reversible which provide good choice as integrators for MD simulations. [5,72-76]

3.1.2. Energy Minimization

The conformations of a macromolecule can be defined as the different arrangements its atoms can adopt in the Cartesian space, when they are repositioned through all their degrees of freedom. An N- dimensional potential energy surface is, thus, created where N are the molecular degrees of freedom. Many factors contribute to the creation of such hyper-surface including angle bending, torsional allowance around rotatable bonds, bond stretching and interatomic contacts. Energy Minimization (EM) algorithms sample the conformational space trying to find that with the lowest potential energy, or the deepest point of the hyper-surface. The reason for this is to optimize the molecular geometry prior

to an MD simulation through elimination of undesired interactions or non-physical contacts.[90,91]

However, the potential energy surface contains multiple minima. The deepest point in the surface is called global minimum; yet it contains multiple local minima or substates. Between minima there are humps or saddles that represent conformations of greater energy that have to be sampled to reach the following minimum. Ideally the global minimum is pursued, but in the practice, it is impossible to sample the whole conformational space in a reasonable amount of time and the arrival to the global minima is never guaranteed. In many cases the algorithm just aspires to a certain threshold of Energy/Forces.[88-92]

3.1.3. Solvation of the system

It is widely accepted that the inclusion of solvent is indispensable to reproduce some properties of soluble biomolecules.[93] In an MD simulation, the solvent can be added as a continuum medium (implicit solvation) that intends to emulate the solvent in energy terms.[94] Another way of including solvent in the simulation is by adding discrete particles of water explicitly. Although the inclusion of explicit waters consumes much more computational power, it has been demonstrated how the solvent may influence internal motions and functionality of macromolecules.[95]

Calculation of all the forces produced in a bulk of solvent can be computationally expensive, therefore must be limited. Simulation in a water box outside of which is vacuum may not reproduce properly the properties of the bulk and cause artefacts or solvent molecules may be lost during the simulation. To avoid these problems MD programs employ what is called “Periodic Boundary Conditions”.[96]

3.1.4. Periodic Boundary Conditions

Periodic Boundary Conditions (PBC) reproduces the simulation box (unit cell) in the three-dimensional space creating an infinite 3D-array. Only the content of one cell is simulated but the effects of the particles interacting with the close particles in neighbouring cells are reproduced through the whole periodic array. If one particle escapes from the box it will enter again through the opposite site side of the cell, so that, the number of particles remains constant throughout the simulation.[97]

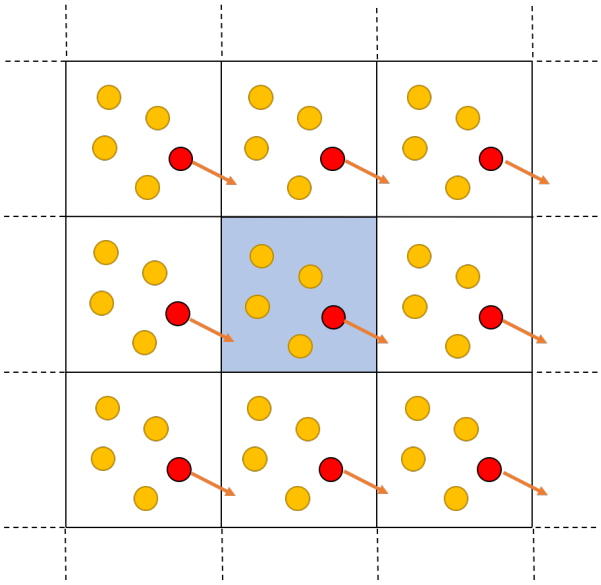

Figure 3.3: *Periodic boundary conditions are shown in two dimensions. Only the content of the central cell is simulated, but its content is reproduced in the periodic array so as to capture those particles leaving the cell.*

3.1.5. Pressure and Temperature Coupling.

MD methods also implement algorithms that maintain and/or modulate pressure and temperature in the simulated system. An ensemble can be defined as a thermodynamic state. MD simulations employ the microcanonical ensemble (NVE). Such ensemble is characterised by a constant number of molecules (N), a constant volume (V) and no energy exchange (E). NVE ensemble corresponds to an adiabatic process where there is not exchange of energy with the exterior, only potential and kinetics energy exchange inside

the system. There are number of reasons why the conditions of the simulation need to be changed: over-heating that causes frictions and calculus errors. Maintaining constant pressure and temperature can emulate the lab and biological conditions.[92,98,99]

In the canonical ensemble or NVT, the temperature is constant during the whole simulation. A thermostat allows the energy to be exchanged, corresponding this to an isothermal process. However, the ensemble that best reproduces lab conditions is called NPT ensemble, since the number of molecules (N), temperature (T) and pressure (P) remain constant. A barostat is included that allows a modulation of the box size to maintain a constant pressure along the run. This last ensemble would then correspond to an isothermal and isobaric process.[90,97]

3.1.6. Coarse- Grained Molecular Dynamics.

Some processes in nature occur above the time-scale that can be reproduced by MD simulations in a reasonable computing time (ns to μs timescale), protein folding or lipid self-assembly, are among these processes . Simplified description of the biological system can be used to accelerate the MD run. The Coarse-grained (CG) models group atoms in coarse particles, so that the degrees of freedom are reduced to speed up the simulation.[100,101]

One of the biological process that can be very computationally expensive, and on the other hand is not necessary to be studied in atomistic detail, is the lipid membrane self-assembly. This process usually occurs in the microsecond timescale. When mixed in solution, amphipathic phospholipids tend to associate in ordered phase in a process driven by the hydrophobic effect. As we are working with membrane proteins of which we do not have previous knowledge of how they interact with the membrane, we follow the process of membrane insertion via self-assembly. In these simulations, the membrane components are allowed to self-assemble in the same system the protein is solvated.[102,103]

3.1.6.1. Martini Force-Field

Martini has been our force field of choice to conduct the CG simulation.[104] Martini groups atoms in clusters of four atoms called "beans". It considers four bean types regarding their physico-chemical properties, namely: Charged (Q), Polar (P), Non-polar (N) and Apolar (A). Each one of those particles has subtypes, so that the properties can better outline the physical profile. For example, the capacity of forming hydrogen bonds is denoted by letters, namely: donor (d), acceptor (a), both (da) or none (0). Or the level of polarity is denoted by numbers, 1-5.[105,106,104]

Martini force field can be implemented in Gromacs, and it is very similar to Gromos force field but the calculations are adapted to the big beans. After the simulation, atomic-scale detail can be obtained by back conversion of CG beans to their atomistic coordinates.[102-104]

3.1.7. Gromacs

We have chosen Gromacs to be our preferred MD engine to perform our simulations. Gromacs implements different atomistic force fields. Gromos force field was designed by the creators of Gromacs and used to conduct atomistic MD simulations in this project.[107,108]

Gromos is an atomistic force field that uses united-atoms particles to describe biomolecular systems, in which non-polar hydrogens are treated as part of the neighbouring heavier atoms.[105, 106]

The particulars options of simulations that Gromacs have to offer and how we carried out our simulations will be explained in the section of methods.

3.2. Experimental

3.2.1. Coarse- Grained Molecular Dynamics (CG-MD)

In order to study the insertion of the proteins in the membrane we carried out five CG-MD simulations, employing the Martini force field.

SIMULATION	*SYSTEM*	*PDB ID*	*LENGTH* (μs)	*DPPC*	*WATERS*
1- GG	GCase	1OGS	1.20	300	5000
2- CG	GCase + GluCer	1OGS	1.20	338	6431
3- CG	CPX	2NSX + 2GTG (pose 5)	1.20	414	8500
4A- CG	Sap-C (closed)	2GTG	1.20	250	4000
4B- CG	Sap-C (open)	2QYP	1.20	250	4000

Table 3.1*: Summary of the CG-MD carried out in this study. Five different systems were inserted into the membrane via self-assembly simulations. They included (1) GCase, (2) GCase bound to its natural substrate, (3) GCase bound to Sap-C and GluCer, and (4) Sap-C.*

The atomistic models were converted using the script "Martinize" obtained from the Martini website. A box of DPPC (Dipalmitoyl Phosphatidyl Choline) lipids was generated. DPPC is the most widely form of Phosphatidyl Choline (PC), the majoritary component of the lysosmal membrane.[109] The optimum numbers of lipids for each system were identified using trial and error. This was followed by 1000 steps of steepest descent energy minimisation. The systems were then solvated and energy minimised until the desired proportion of water/DPPC lipid ratio was obtained. The energy minimisation in those cycles was conducted in two consecutive steps employing the steepest descent and conjugate gradient (1000 cycles) algorithms. Finally, the molecular dynamics simulations were run for 1.2μs with a time step of 0.003 ns, employing Berendsen temperature and

pressure coupling. CG-MD was performed in Gromacs using Martini force field. The simulations were run on the UCL Legion Super Computer cluster using 24 CPUs.

As shown in Table 3.1, in simulations 2 and 4, Glucosylceramide (GluCer) substrate was used. The CG parameters for GluCer were obtained from the Martini website. The substrate was manually positioned in the active site, using atomistic coordinates as a reference. The lipid tails were extended as the parameters obtained from Martini website account for a molecule with smaller acyl tails.

3.2.2. Atomistic MD (AT-MD)

In order to study our system to atomistic detail we carried out 10 atomistic simulations, including complexes of the two proteins and mutants. The simulations are summarised below.

SIMULATION	*SYSTEM*	*a*	*b*	*N.Atoms*	*AT (ns)*
1	GCase	-	IN	20070	500
2a	GCase + GluCer	ACT		22084	1000
2b	GCase + GluCer		IN	22082	1000
3a	CPX	ACT		26540	1000
3b	CPX		IN	26537	1000
4	SAP-C	-	-	13242	500
5a	CPX (N370S)	ACT	-	26536	1000
5b	CPX (N370S)		IN	26536	1000
6a	CPX (L444P)	ACT	-	26535	1000
6b	CPX (L444P)	-	IN	26535	1000

Table 3.2: *Summary of the AT-MD simulations conducted in this project. In some cases the systems were converted to atomistic detail using both (active/helical and inactive/extended) conformations of the protein GCase. i.e. In simulation 2a the active form of GCase has been used for conversion from CG to MD and in simulation 2b the inactive form of GCase has been used for conversion.*

AT Parametrization

Force field compliant topologies were generated using the proteins of the converted models. GluCer was added in those simulations where it was required by aligning the converted models to the docked structure. CG models were converted using the same AT-coordinates that were used to create them. The converted models were, in some cases, used to obtain different conformations/ mutants of GCase by aligning to the desired structures. The AT conversion of Simulation 2-CG was used to obtain the coordinates for Simulation 2a and 2b, Simulation 3-CG was used to obtain the coordinates for Simulations 3a and 3b, Simulations 5a and 5b and Simulations 6a and 6b.

AT-MD

The models were solvated using Single Point Charge water (SPC), the default solvent in Gromacs. They were energy minimised using 5000 steps of steepest descent method. Counter ions were also added to neutralise the systems. A second round of energy minimisation cycle was conducted employing an additional 5000 steps of the steepest descent method.

Two rounds of equilibration were carried out: Firstly, 0.1 ns of NVT equilibration with time steps of 0.002 ns, using V-rescale algorithm for temperature coupling. The temperature was coupled separately for protein/ complexes, lipids and solvent at 323 K and using a time constant for coupling of 0.1 ps. This was followed by 1 ns of NPT equilibration with a time step of 0.002 ns, using Nose- Hoover temperature coupling and Parrinello-Rahman for pressure coupling. The temperature was coupled separately for protein/ complexes, lipids and solvent at 323 K and using a time constant for coupling of 0.5 ps. The pressure of the system was coupled semi-isotropically using Berendsen algorithm at 1 bar, a compressibility of 4.5e-05 and a time constant for coupling of 5.0 ps.

The production run was carried out for 500/1000 ns without any restraints with a time-step of 0.002. A cut-off of 12 Å was chosen for the neighborlist generation and the Coulomb and Lennard-Jones interactions. Particle-Mesh-Ewald summation was chosen for electrostatic interactions. Nose- Hoover was selected for temperature coupling and Parrinello-Rahman for pressure coupling. The temperature was coupled separately for

protein/ complexes, lipids and solvent at 323 K and using a time constant for coupling of 0.5 ps. The pressure of the system was coupled semi-isotropically using Berendsen algorithm at 1 bar, a compressibility of 4.5e-05 and a time constant for coupling of 2.0 ps. For those systems with two proteins or that included the substrate, strong position restraints were applied for energy minimisation runs and soft restraints for equilibration phase simulations. The production run was carried out without any restraints on the system.

Mutants

Mutant proteins were generated using the molecular modelling package ICM-Pro (www.molsoft.com)[110], using the same PDB structures as for the rest of the models. The mutant simulations were set up by converting CG structures of the complex (3-CG) to atomistic detail, using the mutated proteins instead of the wild type. The simulations were detailed in, Table 3.2.

Analysis

The simulation data was analysed in interactive python using the modules: mdtraj[111], numpy[112] and matplotlib[113]. Gromacs analysis suite was also used as an analysis tool. VMD[114] was used for visualising the trajectories. Pymol[115] was used to produce electrostatic surfaces. The figures were made using Pymol[115], ICM-Pro[110] and VMD[114].

3.3.Results.

3.3.1. Coarse- Grained Molecular dynamics simulations.

Self-assembly coarse-grained simulations have been extensively used to study the orientation of proteins in the membrane[105]. The main goal of the coarse grained simulations was to understand how the protein/complexes associated within the membrane.

Between 40-120 ns all the membranes were assembled. To confirm if the self-assembly process has been successful, we calculated the area per lipid in the bilayer. The proteins/complexes were inserted in to the lipid membrane immediately after their formation and remain anchored throughout the course of the simulation. The CG-MD simulation results suggest that the membranes were well formed and equilibrated.

	1-GCASE-GC	*2-GCASE-GLUCER-GC*	*3-CPX-GC*	*4A-SAP-C-CG*	*4B-SAP-C-CG*
AREA PER LIPID (NM^2)	0.645	0.651	0.640	0.653	0.653

Table 3.3: *Summary of the area per lipid in the CG simulations. The reference value is 0.65 +/- 0.05, we can see that all simulations are within the desirable ranges of values.*[116]

The orientation of both proteins anchored to the membrane was consistent with the experiments.[29,73,117] GCase was oriented with the loops at the entrance of the binding site facing the phospholipid membrane. As GluCer anchored in the membrane, it was anticipated that the active site will be facing the membrane. The orientation of Sap-C alone was very similar to that observed in the experimental studies of membrane interaction, explained in Chapter 2 (Figure 2.7). Furthermore, both GCase and Sap-C remained together during the course of the entire simulation run. This further enhanced our confidence in the selection of the GCase-Sap-C model for further studies.

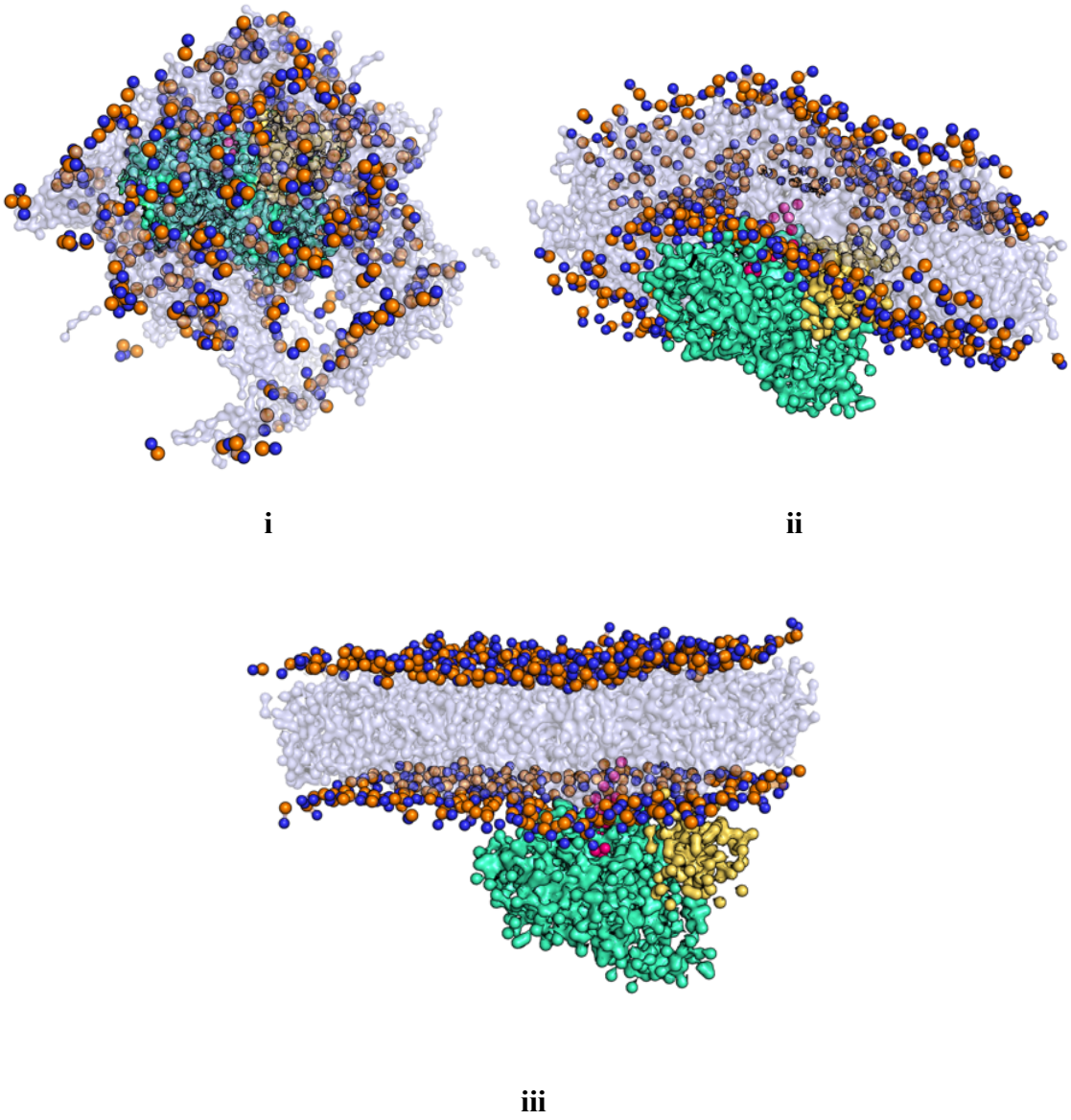

Figure 3.4: *Evolution of the system of GCase, Sap-C and GluCer (Simulation 3-CPX-GC). Snapshots taken at (i) 0 ns, (ii) 30 ns and (iii) 1200 ns. At 30 ns the membrane is not completely formed although the bilayer has taken shape. After approximately 100 ns the membrane is completely formed. The complex of the proteins and the substrate is well anchored till the end of the simulation (1200 ns). DPPC lipids have been coloured in mauve, GCase is been coloured in green, Sap-C in yellow and GluCer in magenta.*

3.3.2. Atomistic Molecular Dynamics Simulations.

3.3.2.1.Wild type Proteins

3.3.2.1.1. General Analysis

The self-assembled coarse-grained systems were re-converted into atomistic detail in order to study the conformational dynamics of the complex in lipid. Table 3.2. provides a list of atomistic simulations carried out.

We started analysing the simulations of wild type GCase in complex with or without Sap-C; namely simulations 1 (GCase-Ext), 2a (GCase-Hel and GluCer), 2b (GCase-Ext and GluCer), 3a (CPX-Hel) and 3b (CPX-Ext), and also Sap-C in simulation 4 (Sap-C). We analysed and compared them to understand how GCase behaved in different complexes and interacted with the different components. We then analysed the simulations of GCase mutants and compared them with each other and with the wild type in order to understand the structural implications of the mutations on the conformational dynamics of GCase.

Conformational drift of a protein or complex was evaluated by measuring the Root-mean square deviation (RMSD) of Cα atoms from the initial structure. We have focussed on GCase in different simulations (and Sap-C in those that contain it) to assess its conformational stability.

Overall, the structures are stable in simulations 2a (GCase-Hel and GluCer), 2b (GCase-Ext and GluCer), 3a (CPX-Hel) and 3b (CPX-Ext), as assessed by RMSD values, shown in Figure 3.5. Equilibration was reached at approximately 250 ns in all the simulations except in 2b (GCase-Ext and GluCer) which does not stabilise until 700 ns, after a period of relaxation of the system. The overall RMSD value or equilibrium value for active GCase when it is simulated without Sap-C in Simulation 2a (GCase-Hel and GluCer) (3.8 Å, 4.1 Å after reaching the equilibrium) is considerably higher than when it is simulated in complex with Sap-C in Simulation 3a (CPX-Hel) (2.7 Å). The difference in RMSD values indicates that the active conformation of GCase is stabilised, when in complex with Sap-C in simulation 3a (CPX-Hel). When we examine inactive GCase, we observe that in simulation 2b (GCase-Ext and GluCer) the equilibrium is not reached until after 700 ns of simulation. When it is simulated in complex with Sap-C in Simulation 3b (CPX-Ext) the

equilibrium value (4.0 Å) was reached earlier in the simulation after approximately 250 ns. The structural changes in the conformations result in greater RMSD values of the extended conformation of GCase in presence of Sap-C, however in absence of Sap-C GCase shows less conformational stability.

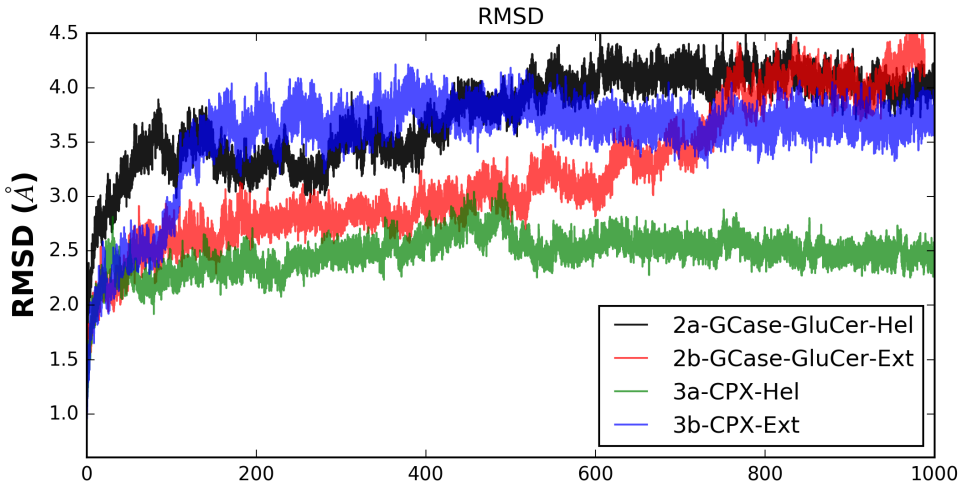

Figure 3.5: *Cα-RMSD values of GCase, plotted as a function of time for simulations 2a, 2b, 3a and 3b.*

In order to better understand conformational flexibility, we measured the RMS Fluctuation for each residue (RMSF). This measure gives a detailed idea of the local flexibility. It is not surprising that the loops on the surface of the protein are more mobile than those within the core. It also indicates that there is not any important conformational drift within the core structure. Results are shown in figure 3.6. Loops at the entrance of the binding site do not show high fluctuation. Loop-1 (311-319) shows similar levels of fluctuation in Simulations 2a (GCase-Hel and GluCer), 3a (CPX-Hel) and 3b (CPX-Ext), between 1 and 2 Å. In Simulation 2b it exhibits higher values of fluctuation. In simulation 2b, Loop-1 extends to helix 7 adopting an even more extended form. Loop-2 (345-351) presents higher values of RMSF in Simulation 2a (GCase-Hel and GluCer) (3Å) and 3b (CPX-Ext) (2Å). In simulation 2a (GCase-Hel and GluCer) Loop-2 gets embedded, while in simulation 3b (CPX-Ext) it gets trapped under Sap-C which prevents the loop to insert into the membrane (Figure 3.15). Loop-3 (395-399) shows a higher fluctuation in Simulation 3b (CPX-Ext),

where it adopts an active form (Figure 3.15), with a peak of 2.2 Å for residue 395. Loop-4 (237-248) peaks in simulation 2a (GCase-Hel and GluCer) and 3a (CPX-Hel) with values of 2 and 3 Å, respectively. Loop-5 does not fluctuate significantly throughout the simulations. Helix 7 which contains the important residue N370 does not show a high fluctuation throughout the simulations, although the value is slightly greater in Simulation 3b (CPX-Ext).Those residues implied in the protein-protein binding (Y11-S12, R44-S45, Q440-D445, S464-S465 and Y487) also show a higher fluctuation throughout Simulation 3b (CPX-Ext).

The other peaks showing greater fluctuation in RMSF plot are mainly surface loops and parts of the protein. For instance, there is an increased fluctuation observed in fragment 296-304, corresponding to helix 5 of the TIM-Barrel as it approaches to the membrane during the course of simulation 2a (GCase-Hel and GluCer) and 2b (GCase-Ext and GluCer). This fluctuation does not occur when GCase in complex with Sap-C. In simulation 2a (GCase-Hel and GluCer), the RMSF peaks at residue 270. This residue is positioned at the edge of helix 4 in the active form of GCase (helical) and interacts with the solvent, whereas in the inactive (extended) form it is connected to the β −sheet 5 via hydrogen bonds, pointing towards the inside of the protein. During the simulation 2a (GCase-Hel and GluCer), residue 270 goes from interacting with the solvent to making an interaction with β -sheet 5, thereby generating a peak in the RMSF graph. During simulation 3a (CPX-Hel), the residue does not make any interactions and the fluctuation is only as a result of being a surface residue. In simulation 3b (CPX-Ext), the residue goes from interacting to non-interacting conformation towards the end of the simulation and in simulation 2b (GCase-Ext and GluCer) it maintains its interactions with the β -sheet 5.

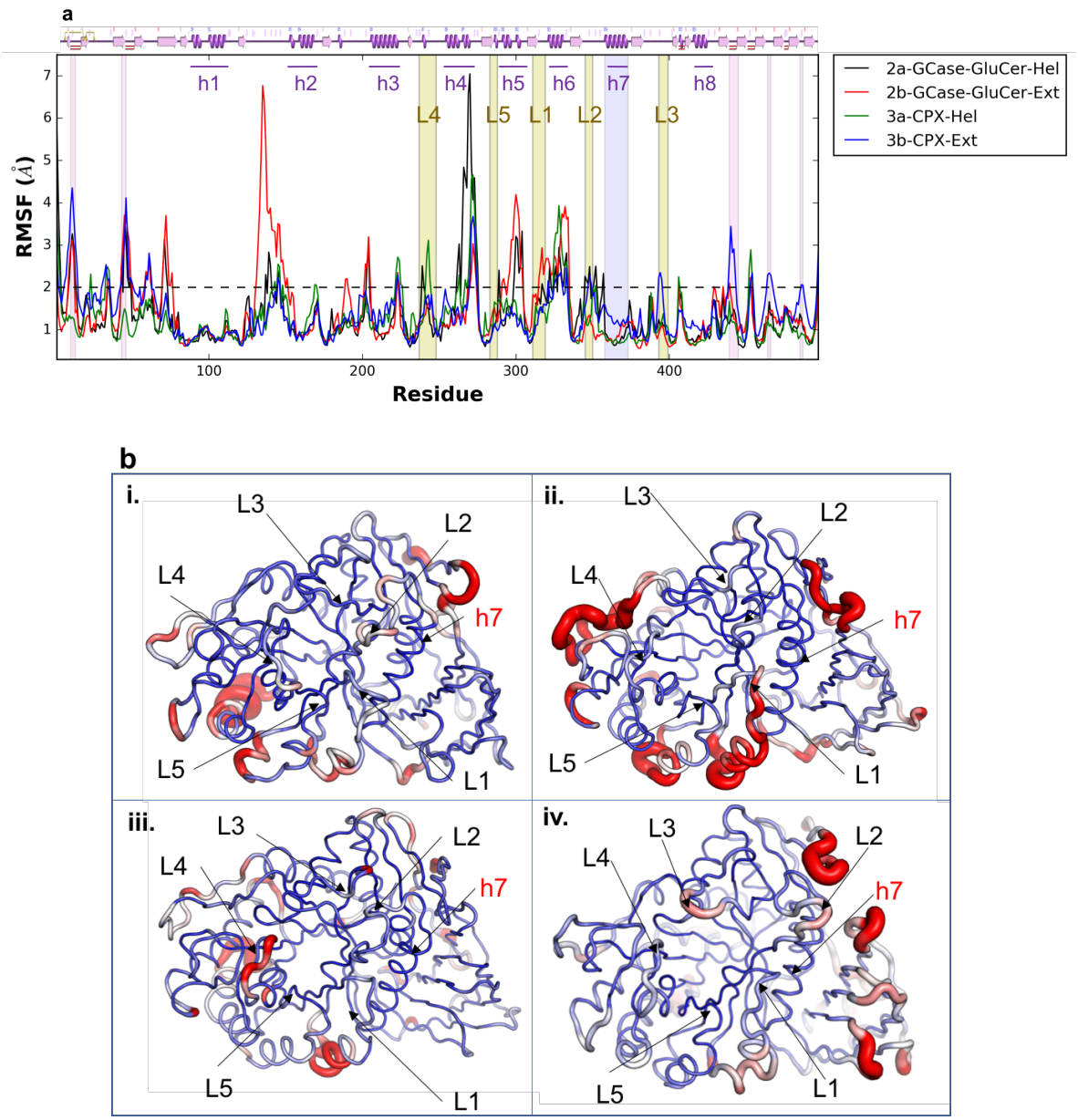

Figure 3.6: *(a) Comparison of RMSF (GCase) as a function of each residue in simulations 2a, 2b, 3a and 3b. Loops-1 to -5 at the entrance of the binding site have been highlighted in yellow and tagged with the label L1 to L5, helix 7 has been highlighted in blue and the protein-protein binding site, other than Loops 1 and 2 and helix 7, has been highlighted in magenta. The secondary structure of the protein can be found on the top of the graph, helices of the TIM barrel have been labelled. (b) RMSF values from simulation translated on the structure of GCase in simulation (i) 2a (GCase-Hel and GluCer), (ii) 2b (GCase-Ext and GluCer),(iii) 3a (CPX-Hel) and (iv) 3b (CPX-Ext)*

Cα RMSD values of Sap-C were also measured when it was present (simulations 3a and 3b) in the model. Sap-C is stable in all three simulations (3a (CPX-Hel), 3b (CPX-Ext) and 4 (Sap-C)). The higher RMSD value was observed in simulation 3b (CPX-Ext) (4.9 Å). Simulation in 3a (CPX-Hel) showed an RMSD value 3.8 Å. The lowest value was observed in Sap-C when it is simulated alone in simulation 4 (Sap-C) (3.4 Å). Equilibration was reached at approximately 200 ns in all the simulations. As shown in Figure 3.7, RMSD value of simulation 4 (Sap-C) undergoes a dramatic increase between 100 and 150 ns. During this time in the said simulation Sap-C anchors to the membrane and adopts a similar orientation and conformation to that observed in simulation 3a (CPX-Hel) and 3b (CPX-Ext) and that mentioned in experimental studies (Figure 3.8).

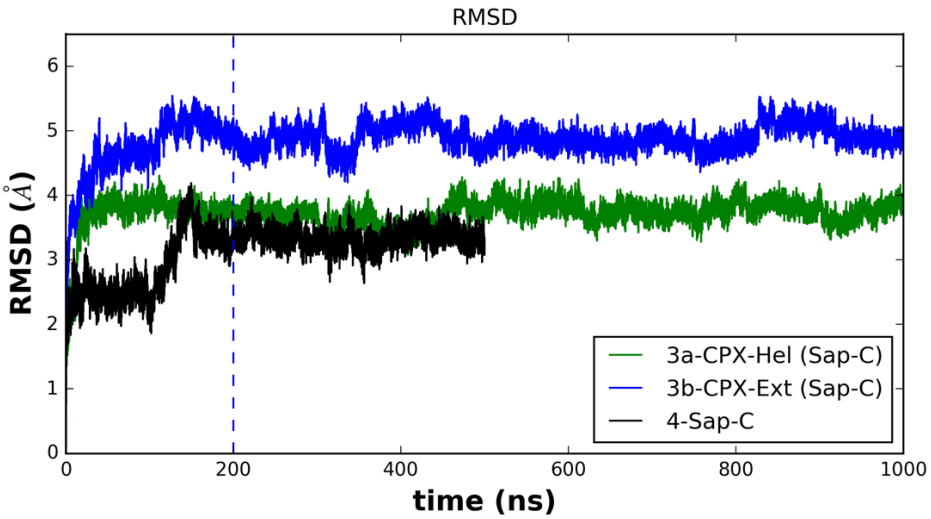

Figure 3.7: *Cα-RMSD values of Sap-C plotted as a function of time for simulations 3a, 3b and 4. The vertical dashed line indicates the time at which the equilibration is reached.*

To understand Sap-C flexibility we calculated the RMSF values during different simulations. Sap-C is a small protein and much simpler in structure than GCase. RMSF values plotted in Figure 3.8, highlight that only the loops that connect helices show a high RMSF value (above 2 Å).

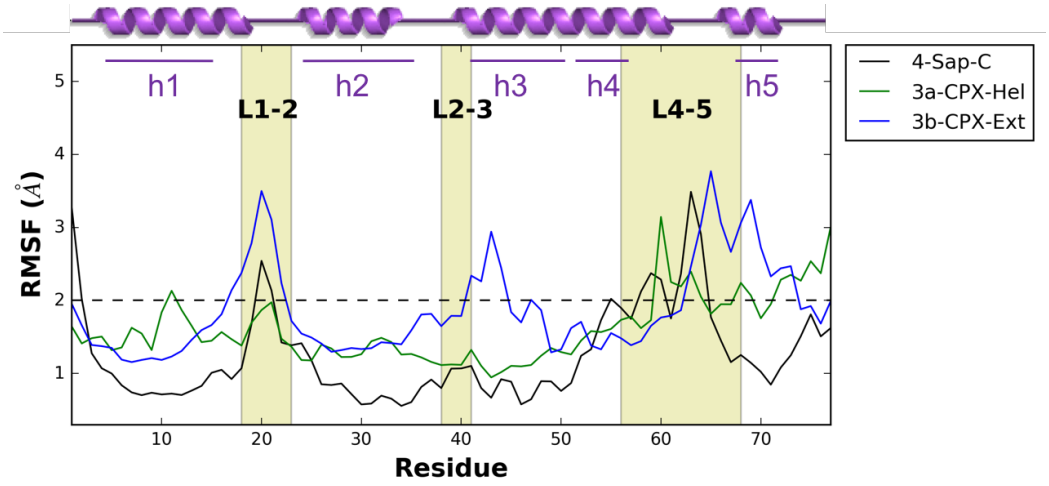

Figure 3.8: *Comparison of RMSF in the facilitator protein Sap-C as a function of each residue in simulations 3a (CPX-Hel), 3b (CPX-Ext) and 4 (Sap-C). Loops joining the helices together have been highlighted in yellow and tagged with the label L1-2, L2-3 to L4-5 referring the helices that join. The secondary structure of the protein can be found on the top of the graph and the helices of the protein have been labelled.*

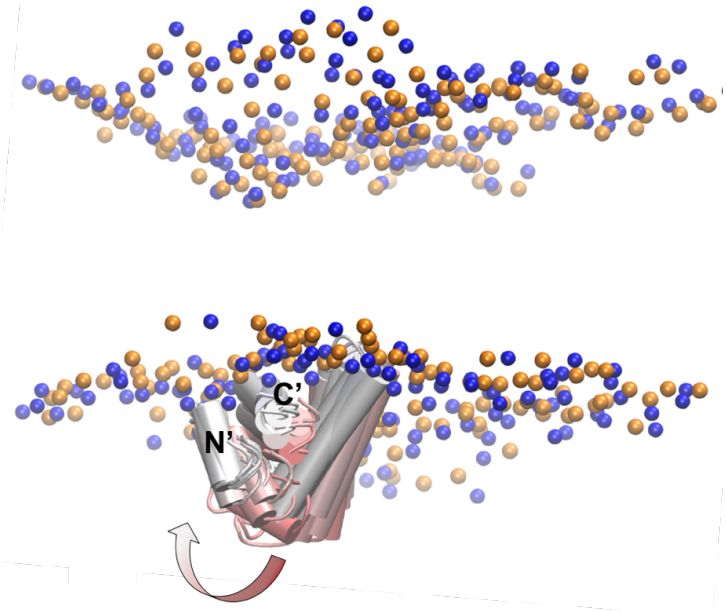

Figure 3.9: *Snapshots of Sap-C in simulation 4, from 50 to 250 ns, period at which the RMSD value of the protein increases dramatically. The protein has been coloured gradually from red (50 ns) to white (250 ns). Helices 1 (C'-terminus) and 5 (N'-terminus) both get embedded in the membrane. Nitrogen and Phosphate atoms of the phospholipids have been depicted as spheres in blue and orange respectively.*

3.3.2.1.2. Membrane Anchoring

It would be useful to count how many and what residues of the protein interact with the membrane as the simulation progresses. We have considered those residues in GCase that lie 4.5 Å away from the membrane to directly interact with it. Table 3.4 summarises these residues that interact with the membrane at 1000 ns. Membrane anchoring gets stronger during the initial equilibration phase, in all the cases. The measure of distance between the centres of mass of the protein and the lipid bilayer as a function of time shows the closeness of the protein to the bilayer. We can observe that the distances between the centres of mass become smaller as the equilibration progresses and remains stable thereafter (Figure 3.5). In simulations of the complex (3a (CPX-Hel) and 3b (CPX-Ext)), the equilibrium distance to the membrane is greater as Sap-C is positioned between GCase and the membrane. It is also worth mentioning that GCase in simulation 3a (CPX-Hel) needs more time to stabilise in the membrane.

SIMULATION	*RESIDUES INTERACTING WITH MEMBRANE*
2A- GCASE- GLUCER	G62, T63, G64, P139, T187, G189, G193, K194, G202, D203, G243, Y244, P245, F246, G250, V294, V295, L296, T297, P299, F316, P319, A320, K321, G325, H328, R329, E349, Q350, S351, V394, Q440, K441, N442, S464, S465
2B- GCASE- GLUCER	G10, Y11, G64, L65, L66, Y135, P139, G189, A190, G193, K194, N200, G202, P253, A292, K293, E300, L314, P319, A320, K321, A322, Q350, S439, Q440, K441, S465, K466
3A- CPX (HEL)	K131, P139, D140, D141, A190, V191, G199, N200, P201, P236, L241, S242, P245, P299, T323, W348, G390, R395, N396
3B-CPX (EXT)	K7, S8, G10, Y11, D127, I130, K131, T132, N188, G189, A190, G193, K194 G199, N200, S237, S242, G243, F316, L317, N350, G389, P391-N396, D405-K408

Table 3.4: *Residues directly interacting with the membrane at 1000 ns. Some differences between active forms of the enzyme and inactive form are observable, as well as how the presence of Sap-C influences the membrane binding.*

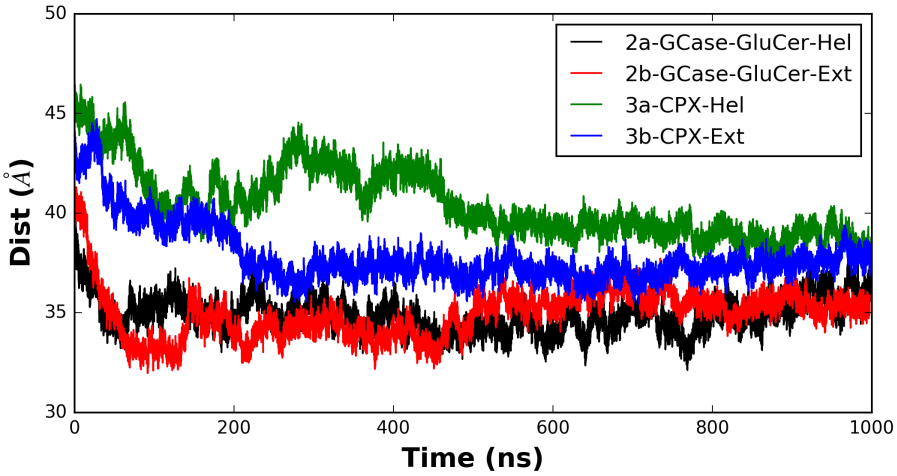

Figure 3.10: *Distance between the centre of mass of GCase and the centre of mass of the lipid membrane.*

3.3.2.1.3. Electrostatic surfaces

The analysis of the electrostatic surfaces of the proteins provides a good understanding of the conformational evolution of the system. Active and inactive GCase show overall similar electrostatic features. At 0 ns, Domain I is positively charged in the regions that face the lipid membrane and less positive away from the bilayer. Domain II, like Domain I, shows more positive character in the proximities of the lipid membrane, although the most positively charged area is a cluster of side chains positioned towards the end of helix 7 and 6 of domain III, and includes positively charged residues: K293 (Helix 5), K321, H328, R329 and H374. Domain III is noticeably negatively charged in the active site, the loops around the binding site are neutral although certain residues provide some electropositivity. The TIM barrel is more positive on the face opposite to the membrane. However, there are some features specific to each simulation. The electrostatic surfaces evolve with the system dynamics. Figure 3.11 and 3.12 shows the evolution of the electrostatic surface of GCase in different simulations, depicted at 0 and 1000 ns of simulation time. The activation or inactivation process that the protein experiences are reflected in the electrostatic surfaces. Furthermore, some of the RMSF characteristics that have been mentioned above are also observed by analysing the electrostatics surfaces.

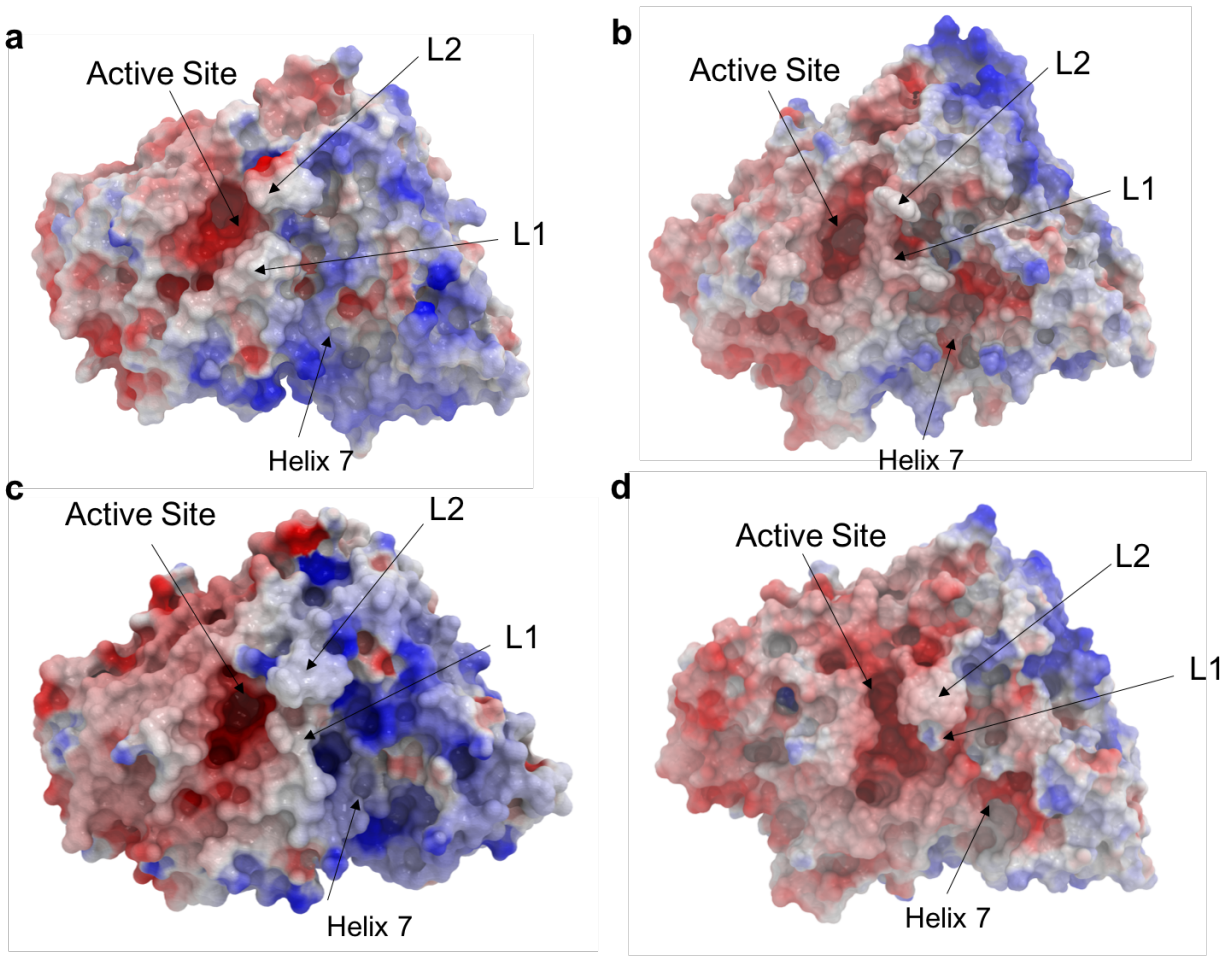

Figure 3.11: *The evolution of the electrostatic surface in simulation 2a (GCase-Hel and GluCer) and 2b (GCase-Ext and GluCer) at 0 and 1000 ns of simulation time. Electrostatic surface of GCase in simulation 2a at (a) 0 ns and (b) 1000 ns. As Loop-1 in simulation 2 loses its helicity, it closes the active site; Loop-4 also progress towards the active site.. Electrostatic surface of GCase in simulation 2b at (c) 0 ns and (d) 1000 ns. It is important to note that Loop-1 extends towards helix 7. The position of Helix 7 has also been highlighted to illustrate the electropositive cluster.*

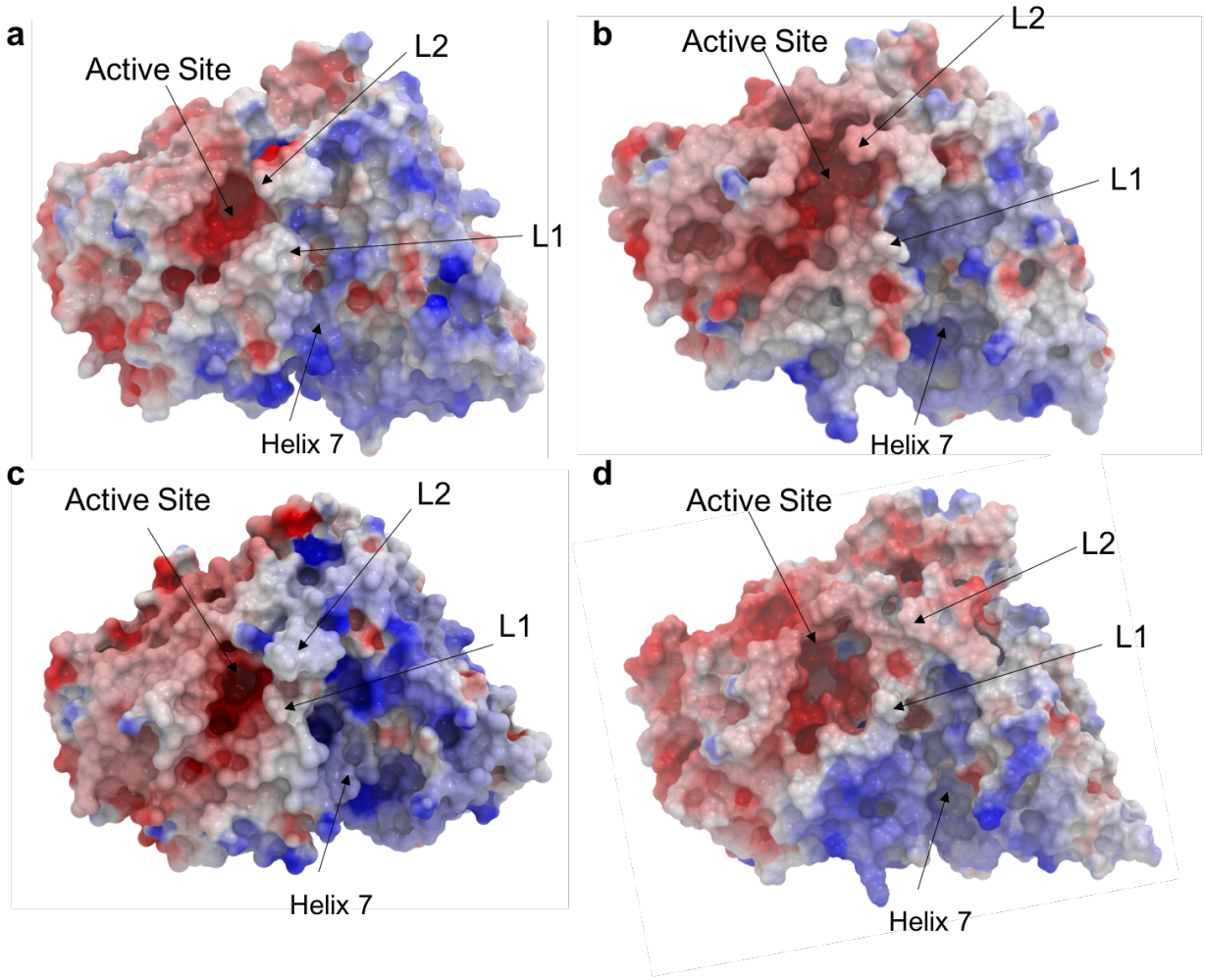

Figure 3.12: *The evolution of the electrostatic surface in simulation 3a (CPX-Hel) and 3b (CPX-Ext). Electrostatic surface of GCase in simulation 3a at (a) 0 ns and (b) 1000 ns. Loop-1 does not change conformation in Simulation 3a as it does in Simulation 3a; Loop-4 progresses towards the active site in simulation 2 although it does not occlude the active site significantly. Electrostatic surface of GCase in simulation 3b at (c) 0 ns and (d) 1000 ns. It is worth noting that Loop-1 and Loop-3 change conformation that leads to widening of the active site. The position of Helix 7 has also been highlighted to illustrate the electropositive cluster.*

3.3.2.1.4. Loop Dynamics

Analysis of the dynamics of the loops at the entrance of the binding site shed light on the activation mechanism of the enzyme. In simulation 2a (GCase-Hel and GluCer) Loop-1 partially loses its helical structure as the simulation progresses. However, when Sap-C is present in simulation 3a (CPX-Hel), the hydrogen bond between D315 of GCase and K33 of Sap-C is maintained over the entire simulation (Fig. 3.13). This hydrogen bond stabilizes the helical conformation of Loop-1. In simulation 3b (CPX-Ext), the side chain of K33 interacts with the backbone atoms of residue L314 and Y373. Although in simulation 3b (CPX-Ext) the helix formation is not complete, there are some observable differences with conformations seen in simulation 2b (when Sap-C is not present) as illustrated in Figure 3.14.

We also observe differences in the evolution of Loop-2 and Loop-3 in the presence and absence of Sap-C. In active state simulation (3a), the side chain of W348 in Loop-2 is oriented towards the outside of the binding site, tucked in a hydrophobic pocket formed by Sap-C. In simulation 3b (CPX-Ext), the side chain of W348 is also trapped under Sap-C. However, in simulation 2b (GCase-Ext and GluCer), the side chain of W348 partially obstructs the entrance of the binding site due to its bulkiness. In simulation 2a (GCase-Hel and GluCer) W348 is embedded in the membrane.

In the inactive state of the enzyme, residue R395 and catalytic residue E340 form a stable hydrogen bond. This hydrogen bond blocks the entrance of the binding site and maintains the catalytic site in the inactive state. This interaction is not observed in simulation 3b (CPX-Ext). During simulation 3b, R395 orients towards the outside of the activation site, ending up in a very similar position as it is in the active state. The interaction between the catalytic residue E340 and residue R395 gets formed during the equilibration time. In the active state, Loop-1 adopts a characteristic helical conformation.

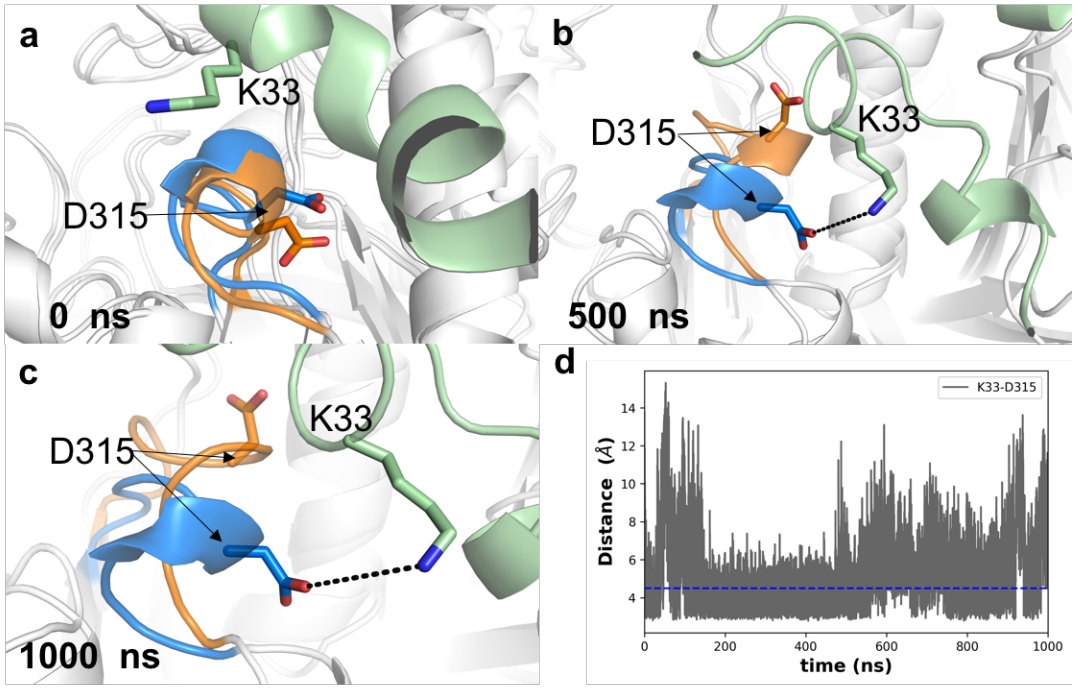

Figure 3.13: *A hydrogen bond between D315 (GCase) and K33 (Sap-C) maintains the helical conformation of Loop-1. Snapshots of conformations extracted from simulation 3a at (a) 0ns, (b) 500ns and (c) 1000ns are illustrated. (d) Minimum distance between K33 and D315 in simulation 3a. GCase has been coloured blue, while Sap-C is coloured green. A comparison of conformations adopted by Loop-1 in simulation 2a (GCase in orange) has also been made at equivalent time and superimpositioned on that of 3a.*

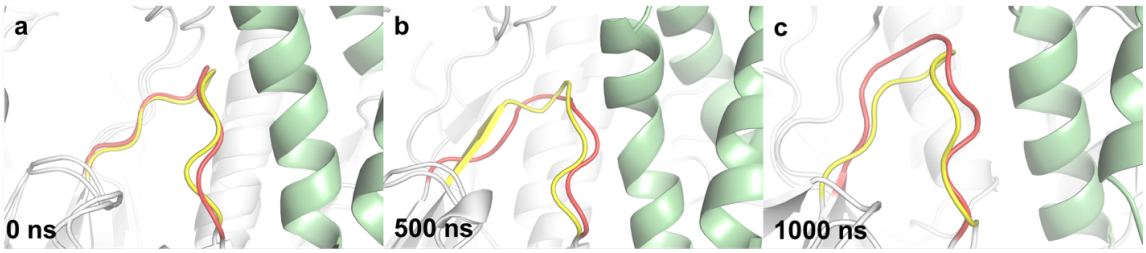

Figure 3.14: *Comparison of conformations adopted by Loop-1 in simulations 2b (red) and 3b (yellow) at (a) 0, (b) 500 and (c) 1000 ns. Loop-1 in simulation 2b extends towards helix 7. The interaction of residue K33 of Sap-C with the neighbouring residues of D315 influences Loop-1 to adopting a helical conformation.*

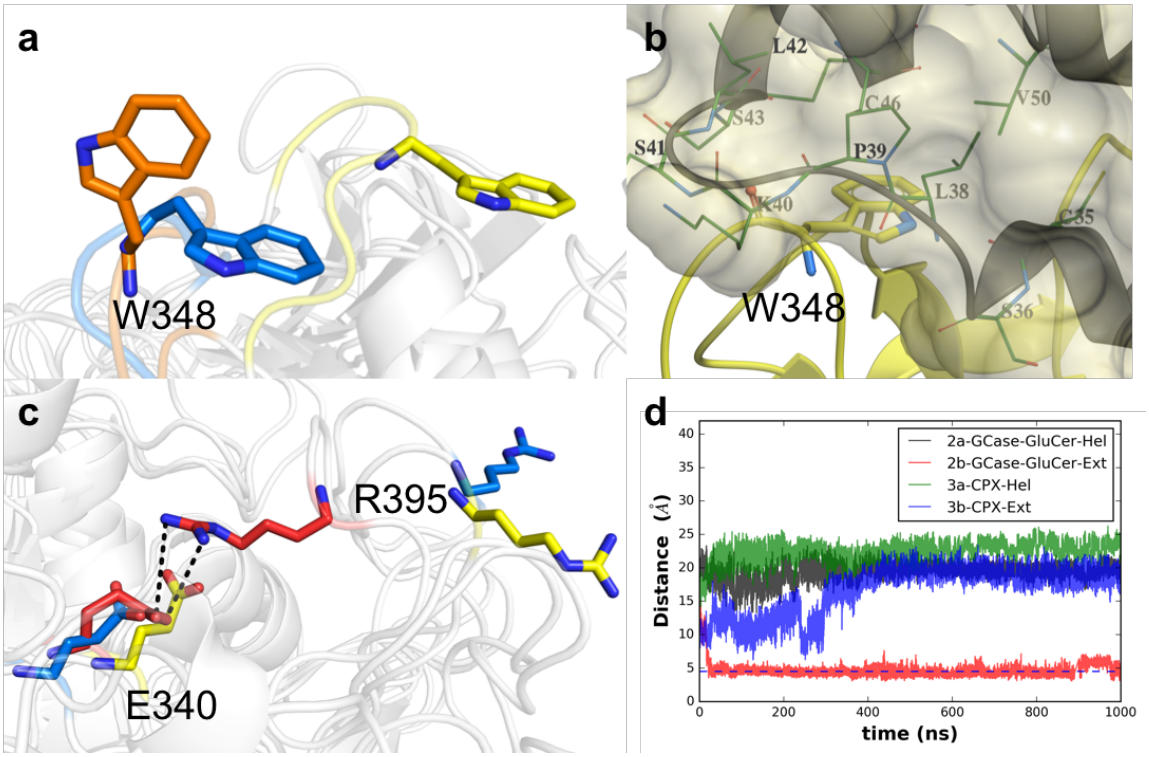

Figure 3.15*: Conformation of the loops at the entrance of the binding site: (a and b) Loop-2 and (c) Loop-3 in simulations 2a (orange), 2b (red), 3a (blue) and 3b (yellow) at 1000 ns.* ***(****b) Snapshot of GCase-Sap-C (green) complex at 1000 ns in simulation 3b. Sap-C stabilizes the active form of the Loop-2, where residue W348 of GCase lies in a hydrophobic pocket formed in Sap-C. (c) Orientation of side chains of R395-E340 in different simulations at 1000ns. (d) Distance between side chains of residues R395 and E340 of GCase in simulation 2a, 2b, 3a and 3b.*

3.3.2.1.5. Interactions in the binding site

Some of the interactions occurring in the active site are of paramount importance to understand the activation process of the enzyme GCase, as well as the implications of some mutations of the protein. Since the interactions made within the binding site have been thoroughly studied and reported in the literature[27], we have not focused on them in this thesis. However, we do need to include some remarks regarding the differences observed at the active sites in different simulations.

There is a high presence of aromatic and hydrophobic residues in the loops at the entrance of the binding site. Firstly, these loops provide an anchor to the phospholipid membrane where hydrophobic effect is the driving force of the interaction.[118] [119] [120] [121] Secondly, aromatic residues are habitually found in the binding sites of glucose and polysaccharides specific proteins. Aromatic residues, specially Tyrosine, Phenylalanine and Tryptophan, have been reported to provide a geometrical complementary surface to the sugar ring of the glucose, the interaction being energetically very favourable.[122] The mentioned residues help in the molecular recognition of the substrate and provide a platform to help it to move through the binding site.

We have found some differences in the evolution of the active sites in all four simulations. In general, the substrate, as it is normal in long MD simulations, moved considerably during the simulation time. As expected, in simulations that start from the active conformation, namely 2a (GCase-Hel and GluCer) and 3a (CPX-Hel), the substrate lasted more time in the docked position. This is probably because in the inactive (Apo) conformation the docking is not optimal since the enzyme is not supposed to interact with the ligand in that conformation.

Along with the catalytic residues E235 and E340, Y313 (Loop-1) plays an important role in guiding the substrate to the catalytic site and stabilising it inside the pocket. This residue is observed to display syncretic behaviour with GluCer in all four simulations (Figure 3.16 to 3.19). Y244 (Loop-4) is another residue that interacts with and prevents the substrate to slip out of the binding site at various times in different simulations.

The presence of Sap-C directly affects the interactions of GluCer in the binding site. It influences a change in the conformation inside the binding site. In simulation 2b (GCase-Ext and GluCer), the interaction between residue E340 and R395 completely blocks the active site (Figure 3.15). GluCer remains positioned over the blocked binding site by making interactions with Y244 (Loop 4) for a while, but that is finally broken and the substrate slips completely out of the binding site.

In simulation 2a (GCase-Hel and GluCer), GluCer goes beyond interacting distance of the catalytic residues (E235 and E340) after approximately 200 ns of simulation, but it remains in the pocket. GluCer establishes interactions with Y313 and Y244. The interaction with Y313 occurs since the beginning of the simulation. Towards the middle of the simulation GluCer interacts with the hydrophobic residues of Loop-4: A238, G239, L241 and L240, and also residues of Loop-5 L283 and L286. GluCer remains between the two loops until the end of the simulation.

In simulation 3a (CPX-Hel), GluCer partially slips out of the active site in the first 150 ns of simulation. The substrate is prevented to completely leave the active site by the interactions made with the aromatic residues W393 (Loop-3) and Y244 (Loop-4), and towards the end of the simulation with the residue F347 (Loop-2) and F316 (Loop-1).

In simulation 3b (CPX-Ext), GluCer leaves the catalytic site in the first 50 ns of simulation. After that, the substrate keeps interacting with residues W393 (Loop-3) and Y244 (Loop-4), and later with residues W393, F397 (Loop-3), Y313 and F316 (Loop-1) until the end of the simulation. Unlike in the inactive form of GCase simulated without Sap-C, the substrate in simulation 3b does not leave the pocket completely.

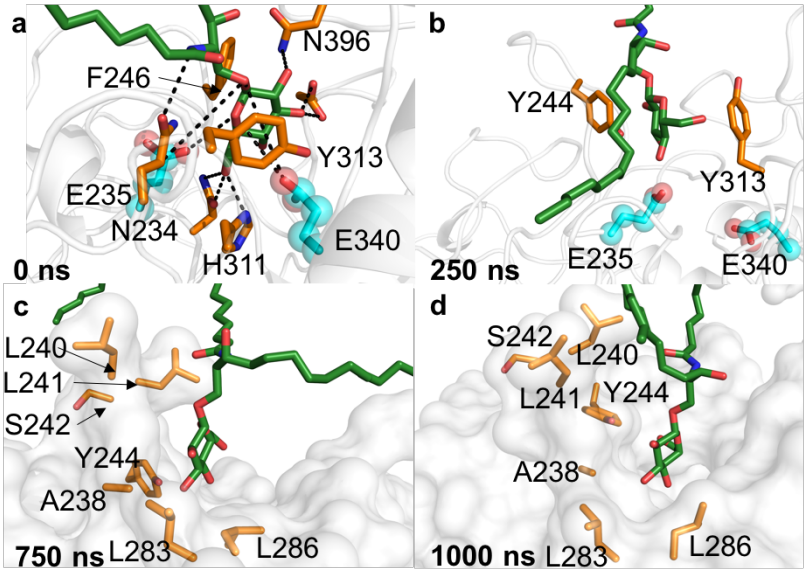

Figure 3.16: *Interactions occurring in the binding site in simulation 2a (GCase-Hel and GluCer) at (a) 0 ns, (b) 250, (c) 750 and (d) 1000 ns. Towards the end of the simulation the substrate is only attached to Loop-4 of GCase. GluCer has been depicted in green, GCase in white with the interacting residues in orange and catalytic residues (E235 and E340) in cyan.*

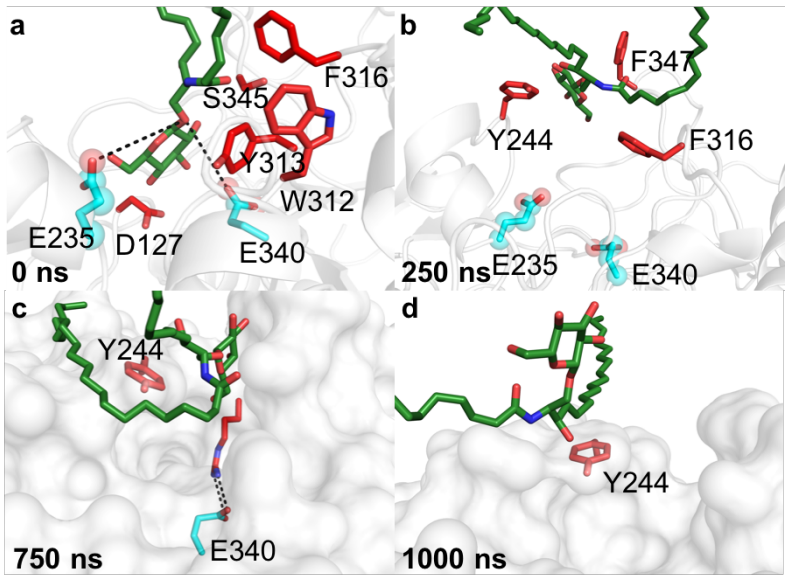

Figure 3.17: *Interactions occurring in the binding site in simulation 2b (GCase-Ext and GluCer) at (a) 0 ns, (b) 250, (c) 750 and (d) 1000 ns. GluCer abandons the binding pocket at the end of the simulation. GluCer has been depicted in green, GCase in white with the interacting residues in red and catalytic residues (E235 and E340) in cyan.*

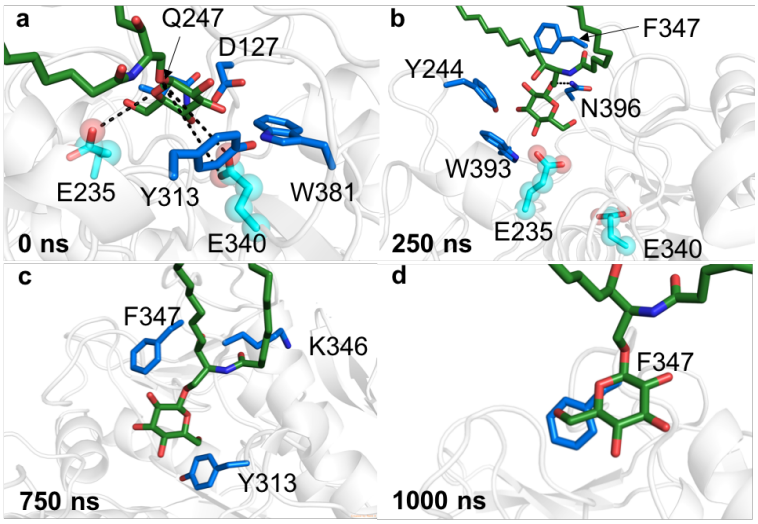

Figure 3.18: *Interactions occurring in the binding site in simulation 3a (CPX-Hel) at (a) 0 ns, (b) 250, (c) 750 and (d) 1000 ns. The substrate makes stable interactions with aromatic residues during the simulation. F347 seems to have an important role maintaining GluCer inside the binding site. GluCer has been depicted in green, GCase in white with the interacting residues in blue and catalytic residues (E235 and E340) in cyan.*

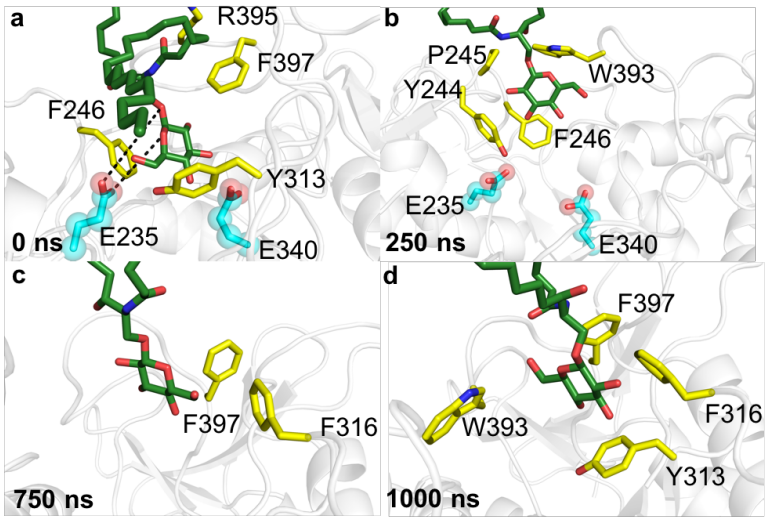

Figure 3.19*: Interactions occurring in the binding site in simulation 3b (CPX-Ext) at (a) 0 ns, (b) 250, (c) 750 and (d) 1000 ns. Similar to the observation in simulation 3a, the substrate makes stable interactions with aromatic residues during the simulations. GluCer has been depicted in green, GCase in white with the interacting residues in yellow and catalytic residues (E235 and E340) in cyan.*

3.3.2.1.6. Protein- protein interactions

There are several Protein-Protein interactions (PPi) that stabilize the GCase-Sap-C complex. The protein-protein interactions occurring in simulation 3a (CPX-Hel) are shown in Figures 3.20 and 3.21 and are summarised in Table 3.5. The protein-protein interface lies between Domain II of GCase and Loop-1 and -2 at the entrance of the binding site, just below helix-7 of the Tim-Barrel. In simulation 3a (CPX-Hel), residue K33 of Sap-C forms stable hydrogen bonds with residue D315 of Loop-1. In Loop-2, we observed one steady hydrogen bond interaction between residues S43 (Sap-C) and W348 (GCase). In helix-7 of the Tim-Barrel (that contains the clinically important residue N370), the interactions between GCase and its facilitator protein are also consistent, as between the residues D29 (Sap-C) and H365 (GCase). Finally, interactions occurring in Domain II of GCase includes those between residue D51 (Sap-C) and R44 and Y487, between S59 (Sap-C) and residues S464 of GCase, between residues S59 (Sap-C) and S464 (GCase) and between residues K25 of Sap-C and N442, D445, D443 (backbone) and L444 (backbone) of GCase.

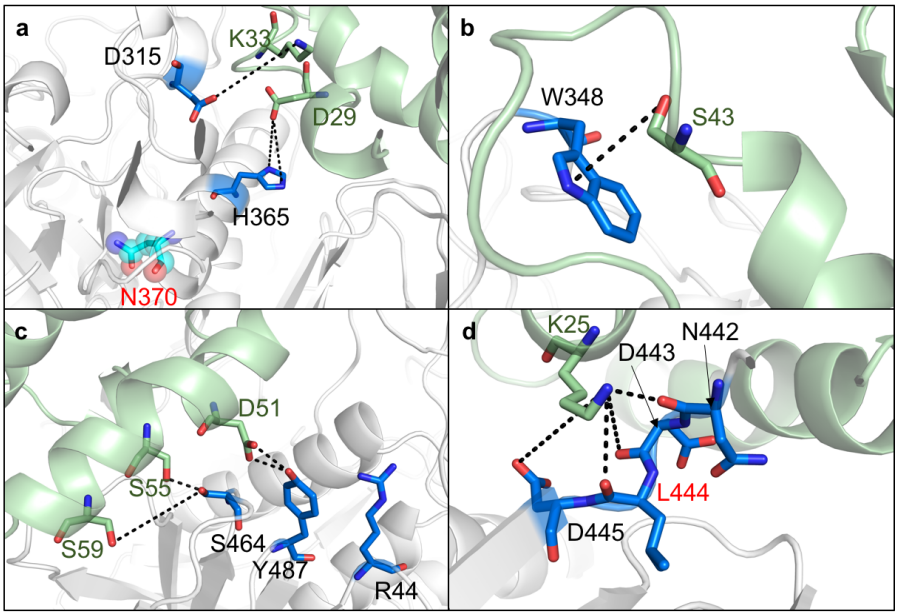

Figure 3.20*: Protein-protein interactions in simulation 3a (CPX-Hel) in (a) Loop-1 and helix 7, (b) Loop-2, (c and d) Domain II.. Sap-C has been coloured in green and GCase has been coloured blue. Important residue N370 has been represented with spheres and coloured in cyan.*

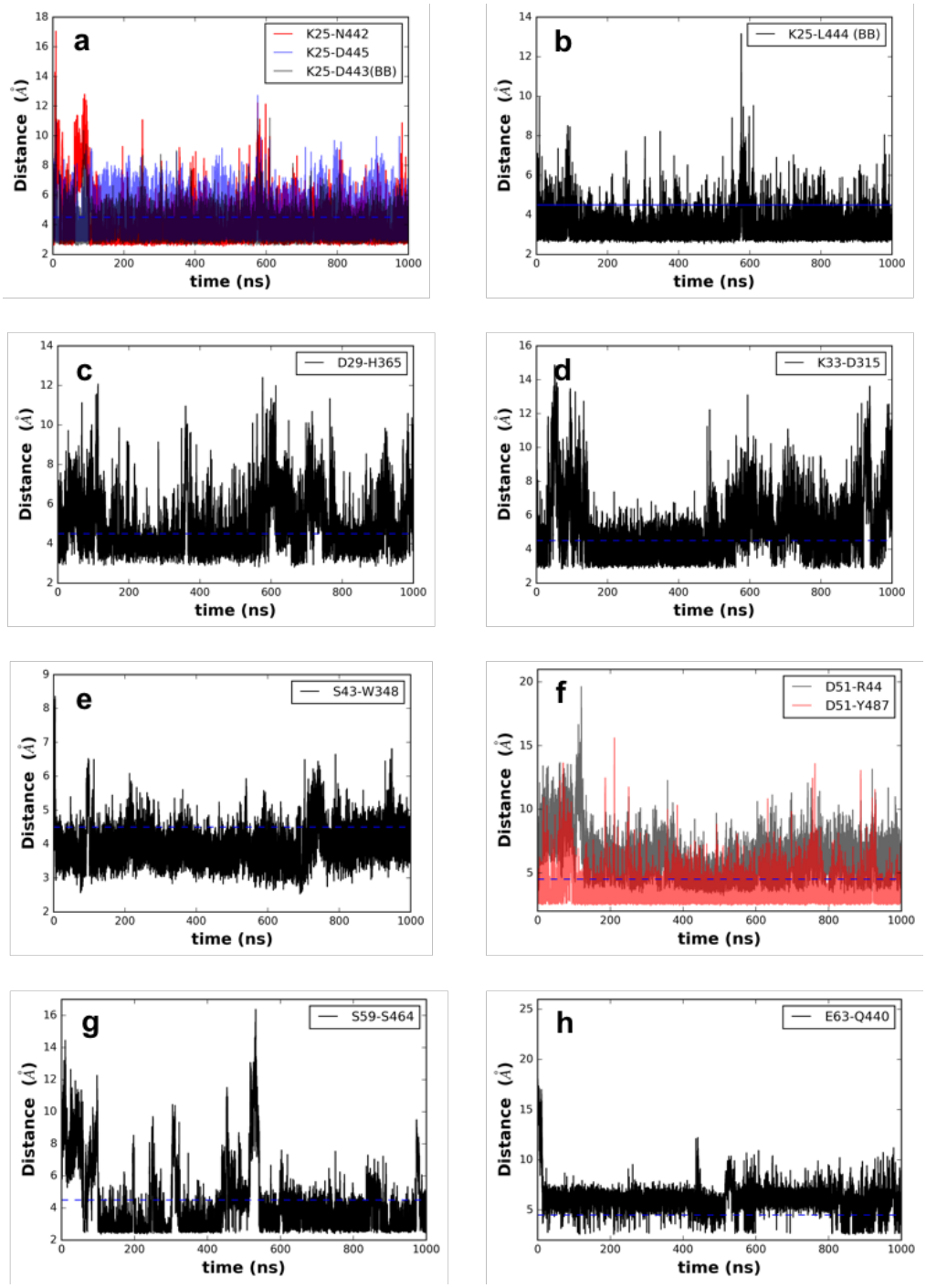

Figure 3.21: *Interactions at the protein-protein interface observed in simulation 3a.*

In simulation 3b (CPX-Ext), residue T23 and K33 of Sap-C forms steady hydrogen bonds with Loop-1 including residue K321 and the backbone of residue L314 respectively. In Loop-2, we observed two very stable hydrogen bond interactions between residues S43 (Sap-C) and E349 (GCase) and between residues S36 (Sap-C) and K346. In helix-7 of the Tim-Barrel, the interactions between GCase and the facilitator protein are also consistent, including that between D32 (Sap-C) and H365 (GCase), between D29 and Y373 (Sap-C and GCase respectively) and between K33 (Sap-C) and Y373. Finally, interactions occurring in Domain II of GCase include those between residues Q47 (Sap-C) and S45 (GCase), between D55 (Sap-C) and R44 and S465, between S55 (Sap-C) and residues S465 and S464 of GCase, and between residues K25 of Sap-C and D443, L444 and D445 of GCase. The protein-protein interactions occurring in simulation 3b (CPX-Ext) are shown in the figures 3.22 and 3.23 and summarised in Table 3.5.

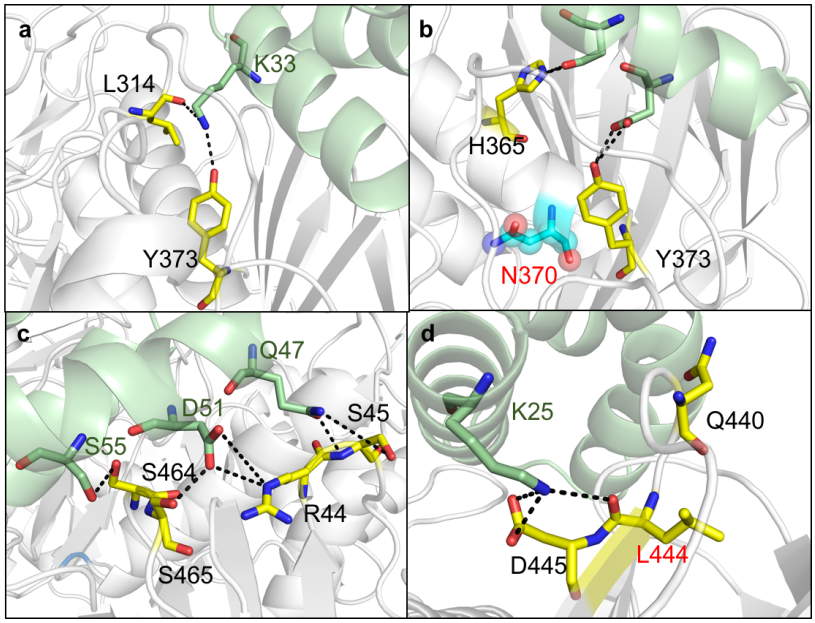

Figure 3.22: *Protein-protein interactions in simulation 3b (CPX-Ext) in (a) Loop-1, (b) Helix 7 and (c and d) Domain II. Sap-C has been coloured in green and GCase has been coloured yellow. Important residue N370 has been represented with spheres and coloured in cyan.*

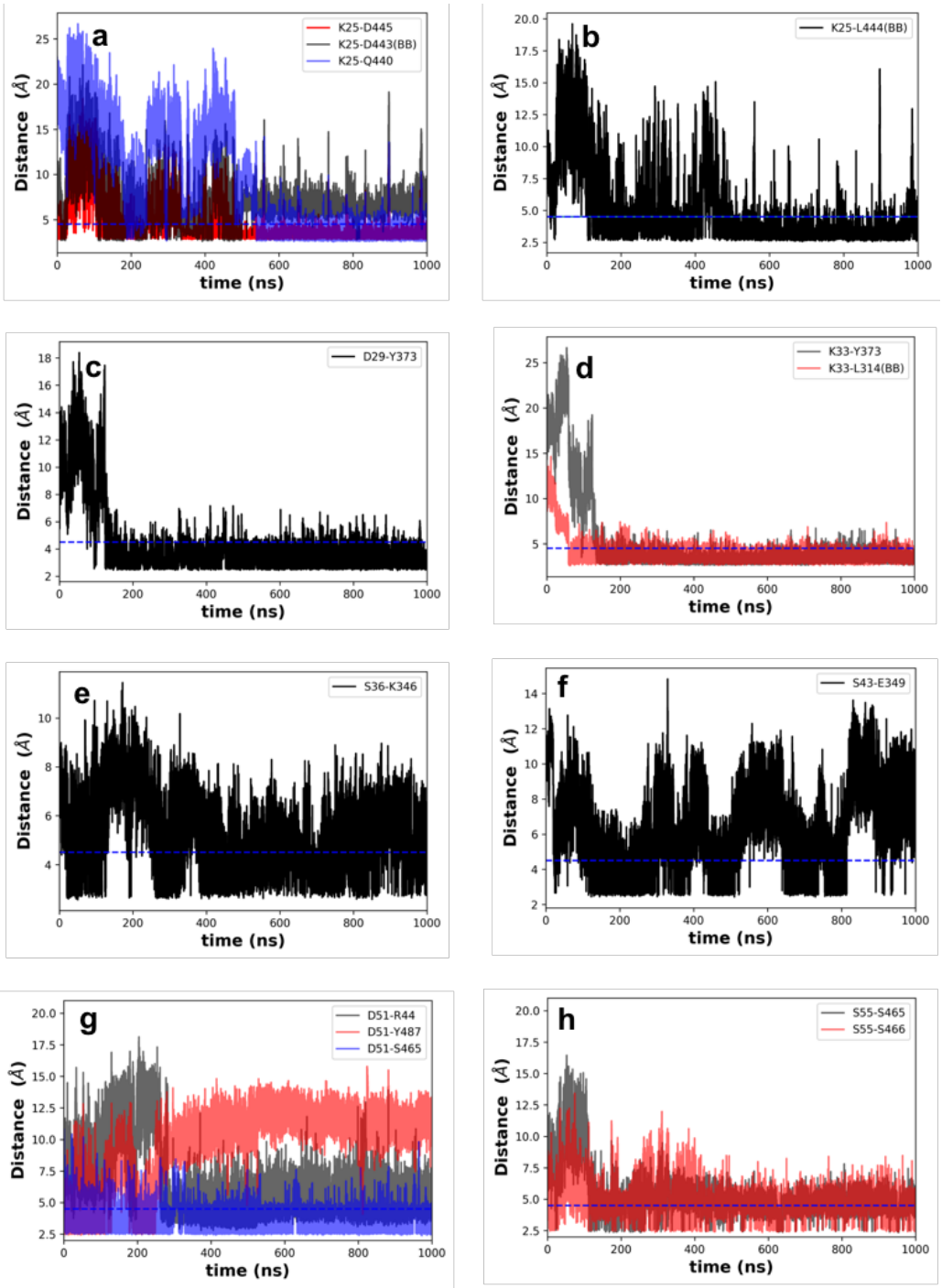

Figure 3.23: *Interactions at the protein-protein interface observed in simulation 3b.*

Sap-C	**3a- CPX (Hel)**	**3b-CPX (Ext)**
T23	-	K321
K25	D445 D443(bb) N442 L444(bb)	D445 D443 L444
E26	-	A320 K321
D29	H365	Y373
D32	-	H365
K33	D315	Y373 L314
S36	-	K346
S43	W348	E349
Q47	D358	S45
D51	Y487 D51(bb)-Y487 R44	R44 (from 250 ns) S465 Y487 (until 300 ns)
S55	Y487 (from 450 ns)	S464 S465
S59	D443 (from 450 ns) S464 (until 400 ns) S465 (until 450 ns)	-
L62(bb)	K441(bb)	K441
E63	Q440 E63(bb)-K441	E63(bb)-K441

Table 3.5: *Summary of protein-protein interaction in simulations 3a and 3b. The abbreviation bb stands for 'backbone'.*

3.3.2.2.Mutant proteins

AT-MD simulations were also performed for two of the most clinically relevant mutants in GCase, namely: N370S and L444P. Both mutant proteins were simulated along with the facilitator protein Sap-C in a membrane environment, and both conformations of GCase were used for each mutant. A total of four simulations were conducted in order to understand the structural implication of these mutations.

3.3.2.2.1. General analysis

To start the analysis, Cα- RMSD of GCase was calculated in all four simulations and compared to the wild type. First, simulations with active conformation (helical) of GCase were analysed, namely 3a (CPX-Hel), 5a (CPX-Hel(N370S)) and 6a (CPX-Hel(L444P)). The RMSD results show an overall conformational stability of the enzyme in the three simulations (Fig. 3.24). The equilibration time in the wild type simulation (3a (CPX-Hel)) was shorter than in its homologue mutants (5a (CPX-Hel(N370S) and 6a (CPX-Hel(L444P)), approximately 100 ns versus 250 ns in the mutants. In simulation 3a (CPX-Hel), the average of the RMSD value from the end of the equilibration is lower (2.4 Å) than in simulation 5a (CPX-Hel(N370S) (3.1 Å) and 6a (CPX-Hel(L444P)) (3.4 Å), indicating a higher conformational stability of the wild type.

Secondly, those simulations containing the inactive (extended form) conformation of GCase were analysed, namely 3b (CPX-Ext), 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)). Once again, the results reveal an overall conformational stability (Fig. 3.25). In simulation 6b (CPX-Ext(L444P)), however, the value of RMSD fluctuates more than in the others. The equilibration time in simulation 3b (CPX-Ext), the wild type, is shorter (~.100 ns) than in the mutants (~ 250 ns). In simulation 3b (CPX-Ext), unlike its active counterpart, the average RMSD value from the end of the equilibration is slightly higher (3.8 Å) than in simulation 5b (3.6 Å) and similar to the average in 6a (3.8 Å), indicating a high conformational flexibility in all three simulations.

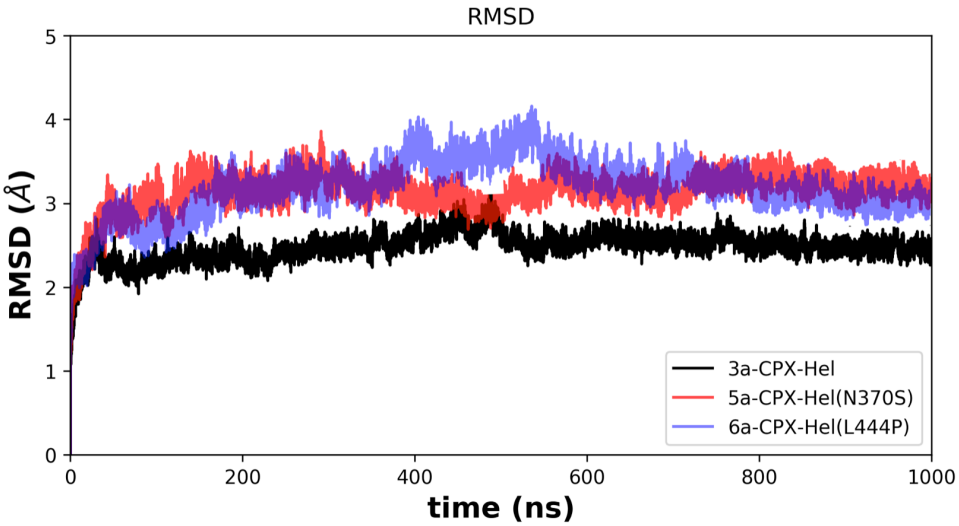

Figure 3.24*: Cα-RMSD values of GCase, plotted as a function of time for simulations 3a, 5a and 6a.*

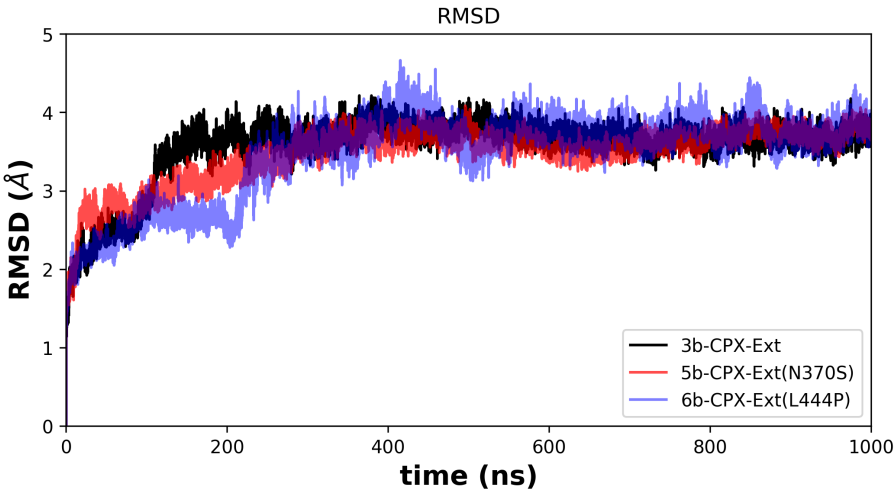

Figure 3.25: *Cα-RMSD values of GCase, plotted as a function of time for simulations 3b, 5b and 6b.*

The RMSF values were also analysed for all four simulations. The simulations of the active form of the enzyme, 3a (CPX-Hel), 5a (CPX-Hel(N370S)) and 6a (CPX-Hel(L444P)) do not exhibit high fluctuation (Figure 3.26). There are only three peaks above 2 Å. Once again, the highest peaks in RMSF values correspond to surface loops, thus indicating a good core stability of the enzyme. Loop-1 does not show high fluctuation in any of the simulations, although it is higher in the mutants than in the wild type (2 Å in 5a, 2.2 Å in 6a, and 1.5 Å in 3a). Loop-2 and Loop-3, do not exhibit high RMSF values in any of the simulations. Loop-4 does not present high fluctuation in any simulation except simulation 3a with an RMSF value of 3 Å. Apart from the mobility of the loops at the entrance of the binding site, helix 6 of the TIM- Barrel (residue 321 to 330) exhibit high RMSF value (7 Å) in simulation 6a, this is because helix 6 deforms towards helix 7, due to the poor docking of the two proteins, losing its helicity. It is also worth noting the high fluctuation of residue N270 in the three simulations. This is primarily due to 270 being a surface residue, which is free to interact with the solvent. Helix 7, which contains the important residue N370, does not show high fluctuation throughout these simulations. On the other hand, those residues implied in the protein-protein binding (Y11-S12, R44-S45, Q440-D445, S464-S465 and Y487) do not show a high fluctuation throughout any of these simulations.

The RMSF values for the extended form of GCase in simulations 3b (CPX-Ext), 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)) were also analysed (Figure 3.27). The core of the enzyme is stable in all three simulations, while the highest peaks correspond to surface loops of the protein. Loop-1 does not show high fluctuation in any of the simulations. Loop-2 exhibit a higher value in simulation 3b with a value of 2.2 Å whereas the mutants showed values lower than 2 Å. Loop-3 again shows a higher value in simulation 3b (2.4 Å) than in its mutant counterparts (1 Å approx.). Loop-4 shows a peak of 2.8 Å in simulation 5b, whereas in the simulation 3b and 6b, it has an RMSF value of 1.9 Å and 1.8 Å respectively. RMSF values of Loop-5 also vary: in simulation 6b it has value of 3 Å, whereas in simulation 3b and 5b, that values are below 2 Å. Helix 5 (residue 269 to 304) and Helix 6 (residue 321 to 330) of the TIM- Barrel display high fluctuation in the two mutant simulations 5b and 6b. Furthermore, an increase in fluctuation is observed for residue 270

in all three simulations. Residue 270 goes from interacting with β-sheet 5 to that which interacts with the solvent. As in simulations 5a and 6a, helix 7 does not show a high fluctuation throughout simulations 5b and 6b. Those residues implied in the protein-protein binding (Y11-S12, R44-S45, Q440-D445, S464-S465 and Y487) also show a higher fluctuation throughout simulation 3b (CPX-Ext) and 6b (CPX-Ext(L444P)).

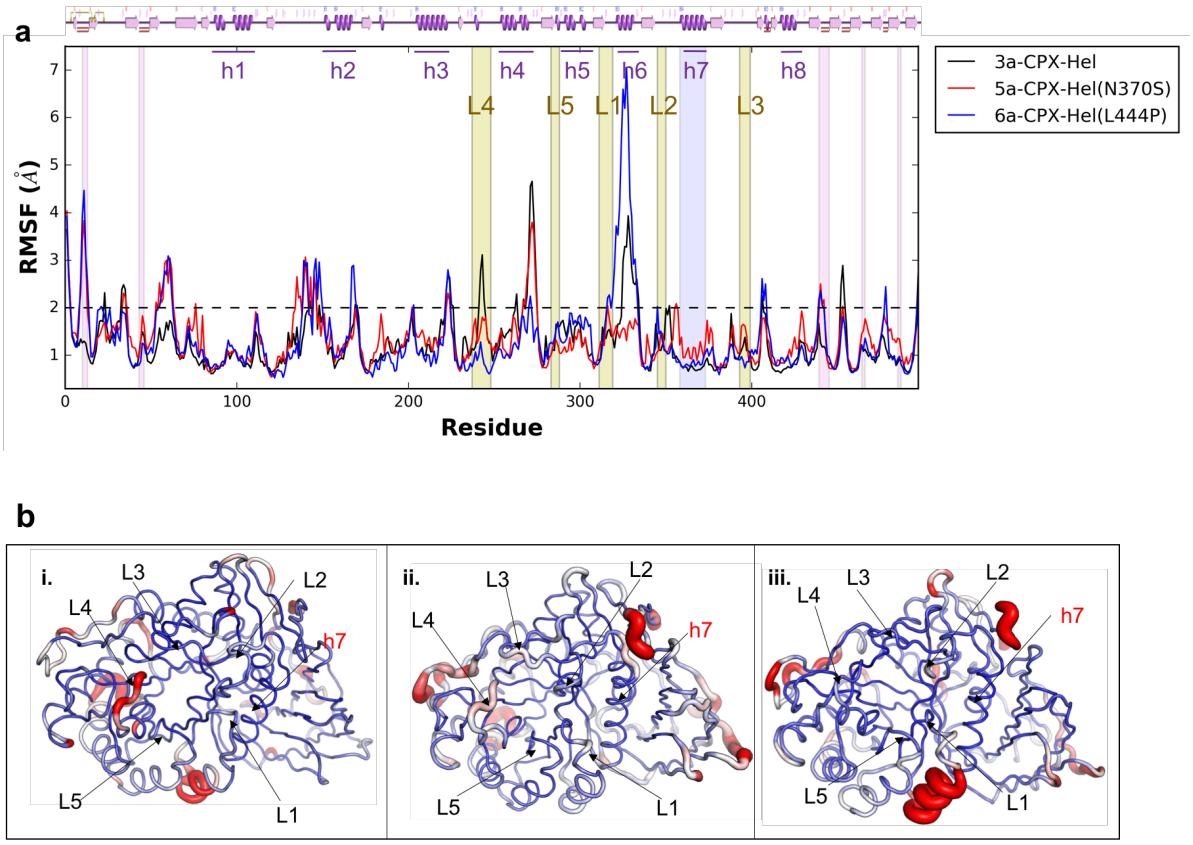

***Figure 3.26**: (a) Comparison of RMSF (GCase) as a function of each residue in simulations 3a (CPX-Hel), 5a (CPX-Hel(N370S)), and 6a (CPX-Hel(L444P)). Loops-1 to -5 at the entrance of the binding site have been highlighted in yellow and tagged with the label L1 to L5, helix 7 has been highlighted in blue and the protein-protein binding site, other than Loops 1 and 2 and helix 7, has been highlighted in magenta. The secondary structure of the protein can be found on the top of the graph, helices of the TIM barrel have been labelled. (b) RMSF values from simulation (i) 3a (CPX-Hel), (ii) 5a (CPX-Hel(N370S)) and (iii) 6a (CPX-Hel(L444P)).*

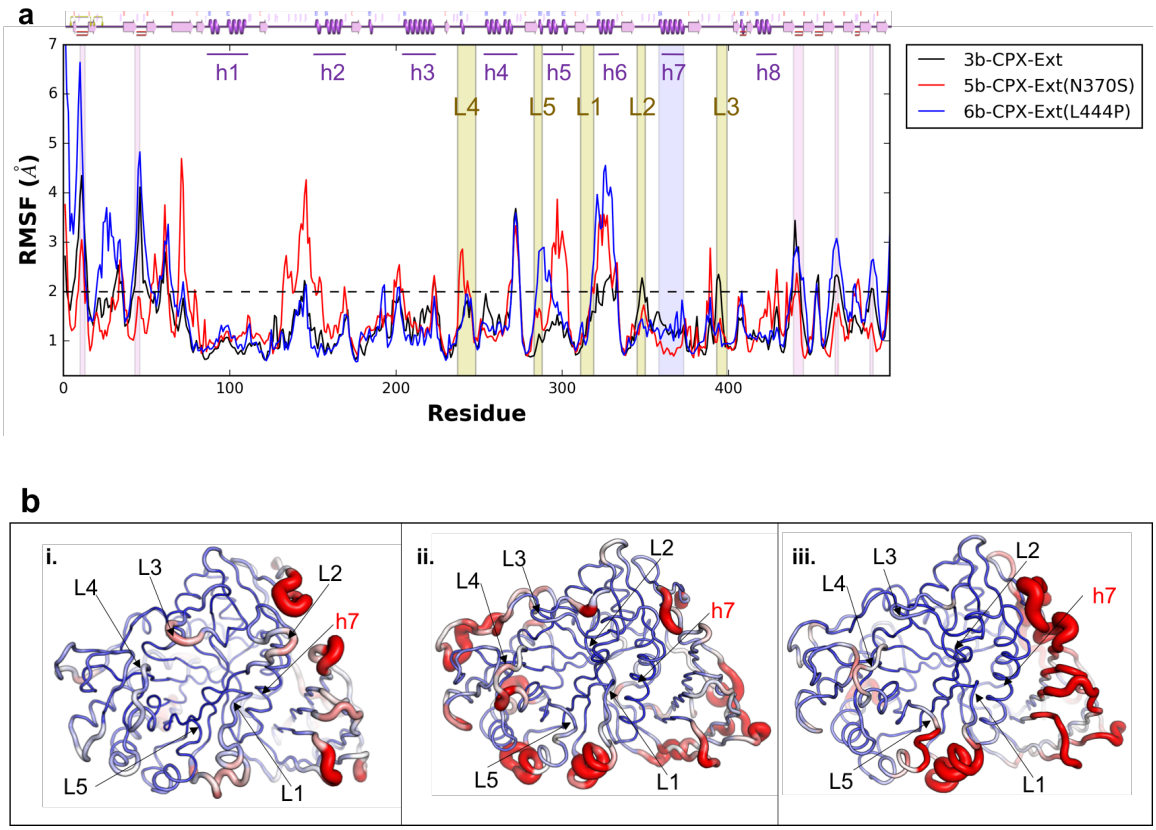

***Figure 3.27**: (a) Comparison of RMSF (GCase) as a function of each residue in simulations 3b (CPX-Ext), 5b (CPX- Ext (N370S)), and 6b (CPX- Ext (L444P)). Loops-1 to -5 at the entrance of the binding site have been highlighted in yellow and tagged with the label L1 to L5, helix 7 has been highlighted in blue and the protein-protein binding site, other than Loops 1 and 2 and helix 7, has been highlighted in magenta. The secondary structure of the protein can be found on the top of the graph, helices of the TIM barrel have been labelled. (b) RMSF values from simulation (i) 3b (CPX- Ext), (ii) 5b (CPX- Ext (N370S)) and (iii) 6b (CPX- Ext (L444P)).*

3.3.2.2.2. Electrostatic surfaces

The electrostatic surfaces of the mutants were analysed throughout the course of the simulations as a way to evaluate the overall dynamics of the mutant proteins. Figures 3.28 and 3.29 show the evolution of the electrostatic surfaces in simulation 5a (CPX-Hel(N370S)) and 6a (CPX-Hel(L444P)) and in simulation 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)), respectively.

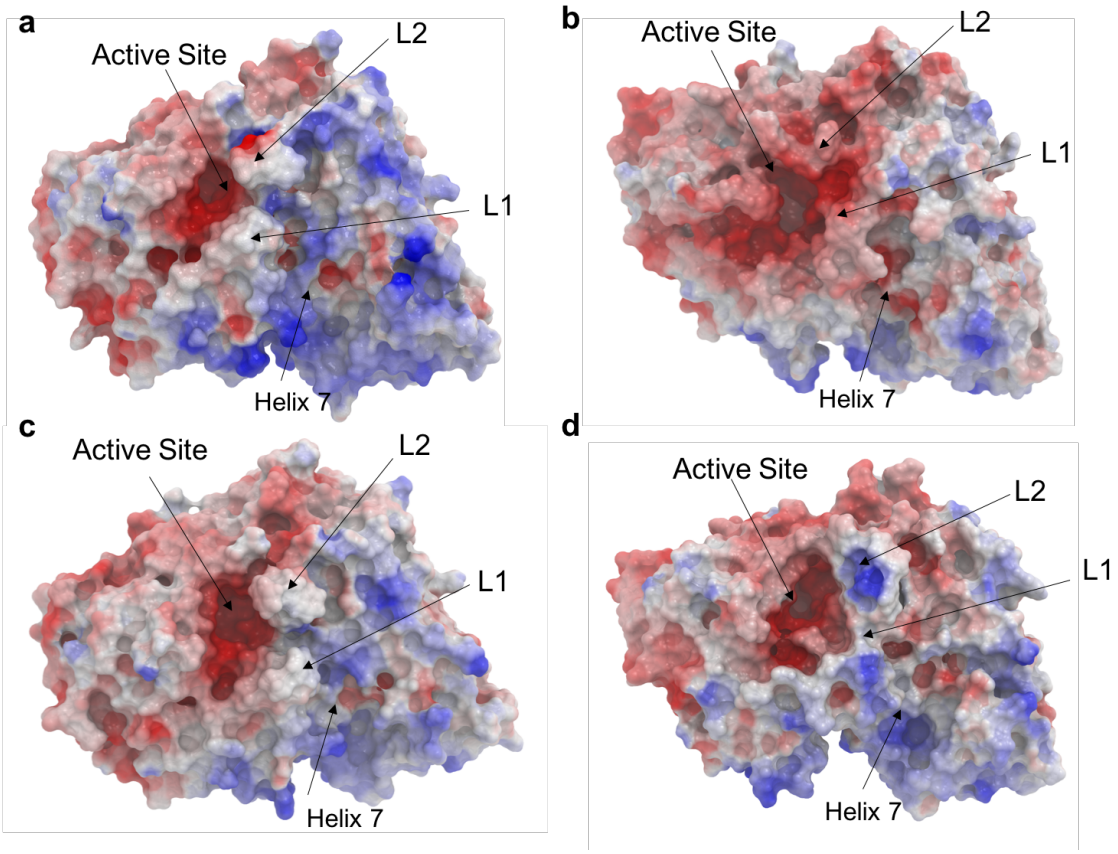

Figure 3.28: *The evolution of the electrostatic surface in both simulation 5a (CPX-Hel(N370S)) and 6a (CPX-Hel(L444P)) is depicted at 0 and 1000 ns of the simulation time. Electrostatic surface of GCase in simulation 5a at (a) 0 ns and (b) 1000 ns. Electrostatic surface of GCase in simulation 6a at (c) 0 ns and (d) 1000 ns. Loop-1, Loop-2 and Helix 7 have been tagged.*

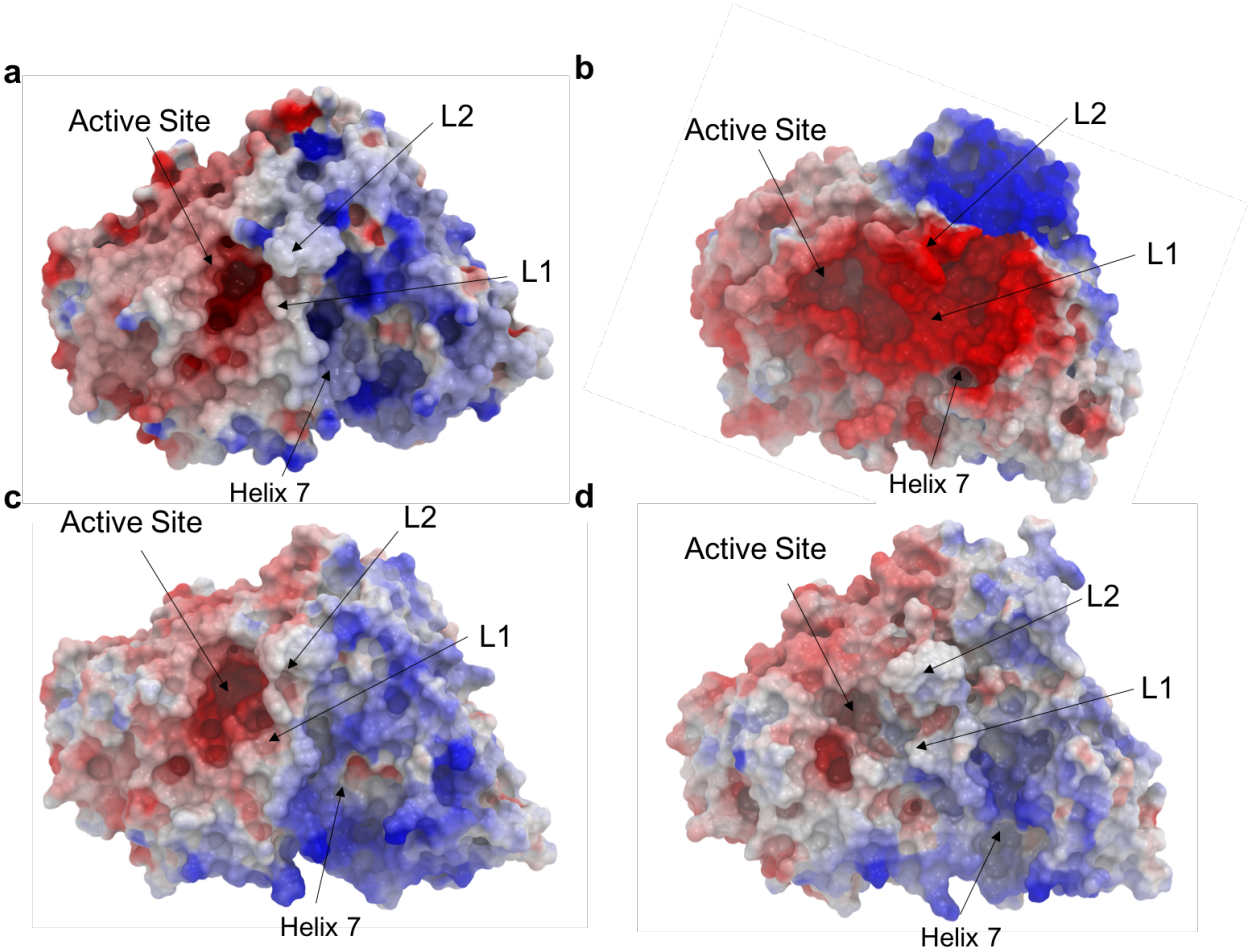

Figure 3.29: *The evolution of the electrostatic surface in both simulation 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)) is depicted at 0 and 1000 ns of the simulation time. Electrostatic surface of GCase in simulation 5b at (a) 0 ns and (b) 1000 ns. Electrostatic surface of GCase in simulation 6b at (c) 0 ns and (c) 1000 ns. Loop-1, Loop-2 and Helix 7 have been tagged.*

3.3.2.2.3. Loops Dynamics

The mutant protein-Sap-C complexes are unstable. This instability affects the dynamic evolution of GCase over the course of the simulation.

The simulations of the mutated protein confirm that loop dynamics is affected by the mutations. In the first 300 ns of simulation 5a (CPX-Hel(N370S)), the helical conformation of Loop-1 is lost in the mutant. This is consistent with when GCase is simulated alone (simulation 2a (GCase-Hel and GluCer)). The helicity of Loop-1 is however partly recovered when interactions between K33 (Sap-C) and the side chain of D315 (GCase) is made during the second half of the simulation (Fig. 3.30). In simulation 6a (CPX-Hel(L444P)), Loop-1 does not lose it helical form during the simulation although the helix gets deformed and moves towards the Loop-2. The helicity is maintained because the formation of a hydrogen bond between D315 and K33 of Sap-C and due to additional interactions with H365 an S366 in helix 7. In the mutant simulations, the bad coupling between the two proteins leaves the Loop-2 free, unlike in the wild type simulations where Loop-2 remained tucked under Sap-C. The evolution of Loop-3 is also different in mutants: while in the wild type, residue R395 is oriented towards the outside of the active site, in simulation 5a (CPX-Hel(N370S)) it is oriented towards the inside forming a hydrogen bond with residue S350 of Loop-2. In simulation 6a (CPX-Hel(L444P)), side chain of residue R395 is still pointing towards the exterior of the binging pocket.

In the simulations of the extended form (inactive) of GCase, the evolution of the loops in the mutants and the wild type highlights some important differences (Fig. 3.31). In simulations 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)), Loop-1 extends towards helix 7. Similar to the helical conformation in simulation 5b and 6b, residue W348 does not remain consistently tucked under Sap-C as it does in simulation 3b, the wild type. In simulations 5b and 6b Loop-3 adopts a closed conformation with residue R395 interacting with the catalytic residue E340. Such an interaction completely obstructs the binding site similar to that observed in simulation 2b (GCase-Ext and GluCer).

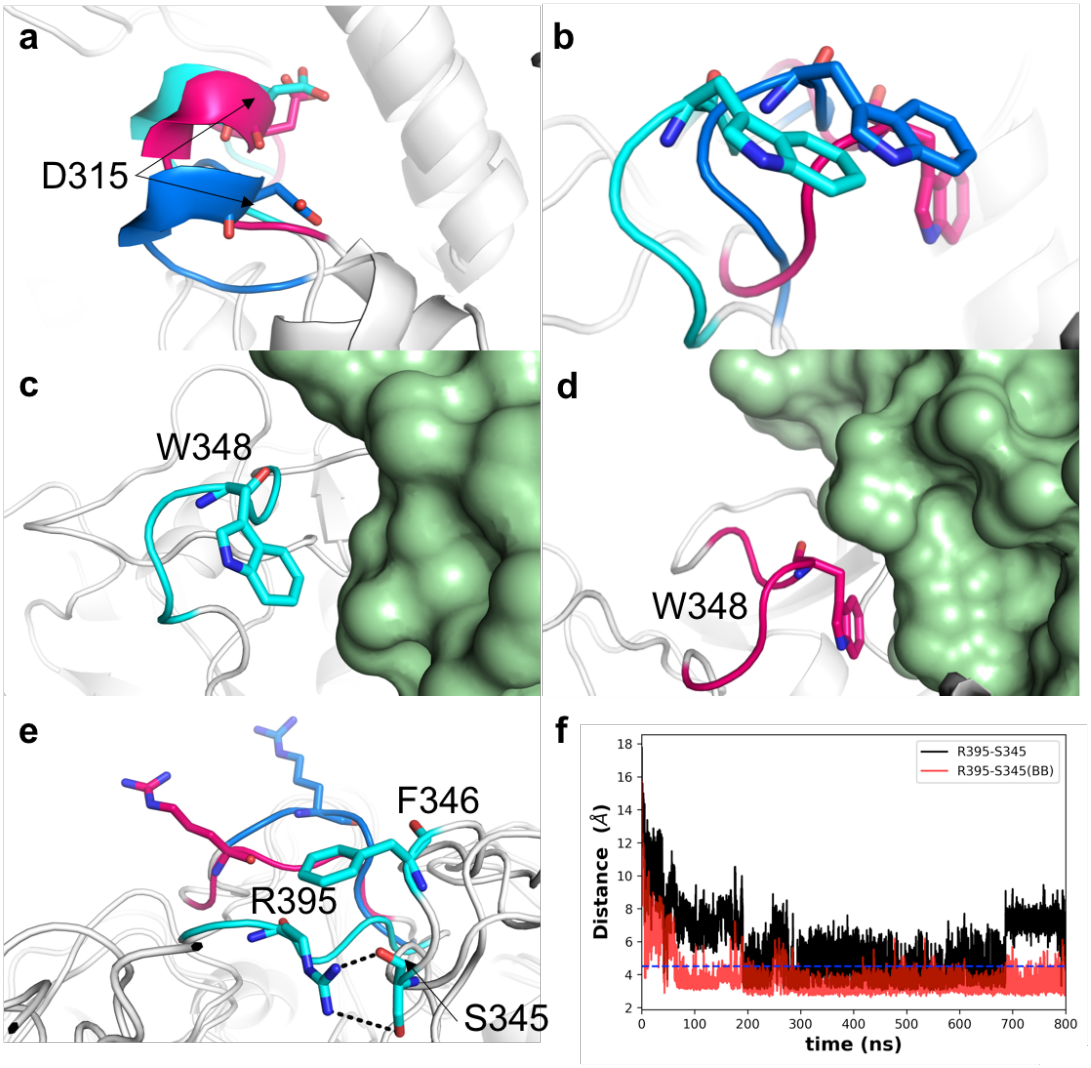

Figure 3.30: *Dynamic evolution of the loops at the entrance of the binding site in the mutant simulations 3a (CPX-Hel), 5a (CPX-Hel(N370S)), and 6a (CPX-Hel(L444P)). Snapshot taken at 1000 ns of simulation time; GCase has been depicted in different colours depending on the simulation: 3a in blue, 5a in hot pink and 6a in cyan and represented as a cartoon, Sap-C has been coloured in green and represented as a surface. (a) Loop-1, (b-d) Loop-2, (e) Loop-3 and (f) distance between residues R395 and S345 in simulation 6a. Loop-1 maintains the helical conformation due to the influence of Sap-C. Further, due to the instability of the protein-protein binding, W348 (Loop-2) does not remain inserted in the hydrophobic pocket in the mutants. Finally, Loop-3 closes towards the binding site in simulation 6a.*

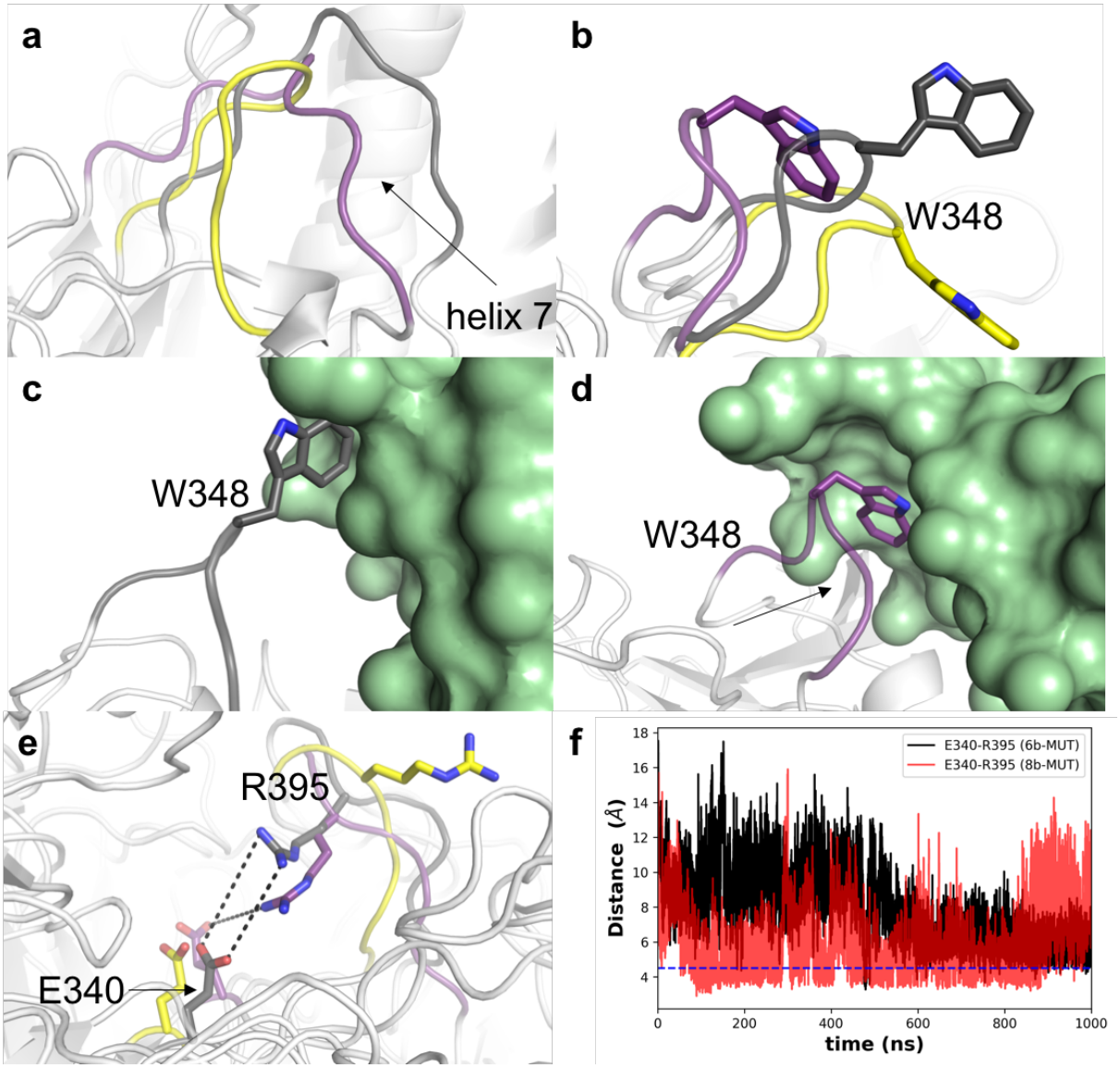

Figure 3.31: *Dynamic evolution of the loops at the entrance of the active site in simulations 3b (CPX-Ext), 5b (CPX- Ext (N370S)), and 6b (CPX- Ext (L444P)). Snapshot taken at 1000 ns of simulation time; GCase has been coloured in yellow (3b), dark grey (5b) and purple (6b), Sap-C has been coloured in Green and GCase has been depicted as a cartoon and Sap-C as a surface. (a) Loop-1, (b-d) Loop-2, (e) Loop-3 and (f) distance between residues R395 and catalytic residue E340 in simulation 6b. Loop-1 extends towards helix-7 in the mutant. Also, the poor binding between the two proteins prevents residue W348 from occupying the hydrophobic pocket in Sap-C. Furthermore, Loop-3 adopts a closed conformation in the mutants.*

3.3.2.2.4. Interactions in the binding site.

Mutations in GCase directly affect the dynamics of the loops and hence the interaction occurring in the binding site. Whereas in the wild type complexes 4a (CPX-Hel) and 4b (CPX-Ext) the substrate remains inside the pocket during the whole simulation time, in their mutant counterparts the substrate leaves the binding pocket along the simulation. We next summarise the most important interactions taking place in the binding site of the four mutants.

In simulation 5a (CPX-Hel(N370S)), GluCer slips out of the catalytic site in the first 50 ns of simulation, but it remains inside the pocket thanks to the interaction with residue Y244 in Loop-4 and residue N396 in Loop-3, as it can be observed in Figure 3.32. Towards the ends of the simulation GluCer also interacts with residue S345 in Loop-2.

In simulation 5b (CPX-Ext(N370S)), the substrate goes completely out of the binding pocket after about 150 ns of simulation. From this point only the acyl tails of the ligand, and momentarily the head, are in contact with the loops of the binding site including some of the residues of Loop-4 like L241 and L240, these interactions are shown in Figure 3.32.

During the first 100 ns of simulation 6a (CPX-Hel(L444P)), GluCer leaves the catalytic site but still remains inside the pocket by interacting with some residues of Loop-2 (F347), Loop-3 (V394, R395 and N396) and Loop-4 (F246) (Fig. 3.33). All these interactions are finally disrupted and the substrate slips completely out of the binding site after 750 ns of the simulation.

Finally, in simulation 6b (CPX-Ext(L444P)) the substrate also abandons the binding pocket. After 400 ns of simulation, GluCer does not have any contact with the residues of the enzyme (Fig. 3.33). Before leaving the active site, the ligand establishes connections with F347 (Loop-2), F397 (Loop-3) and Y244 (Loop-4). Interactions between residues R395 and E340 occlude the entrance of the binding pocket and thereby making it impossible for the substrate to remain inside it.

It is important to note that in all the mutant simulations, except in 5a (CPX-Hel(N370S)), GluCer finishes the simulation outside the binding pocket. This occurs in part because the interaction R395-E340 partly occludes the pocket making impossible for the substrate

come back to the binding site. We have observed this interaction in simulation 2b (GCase-Ext and GluCer).

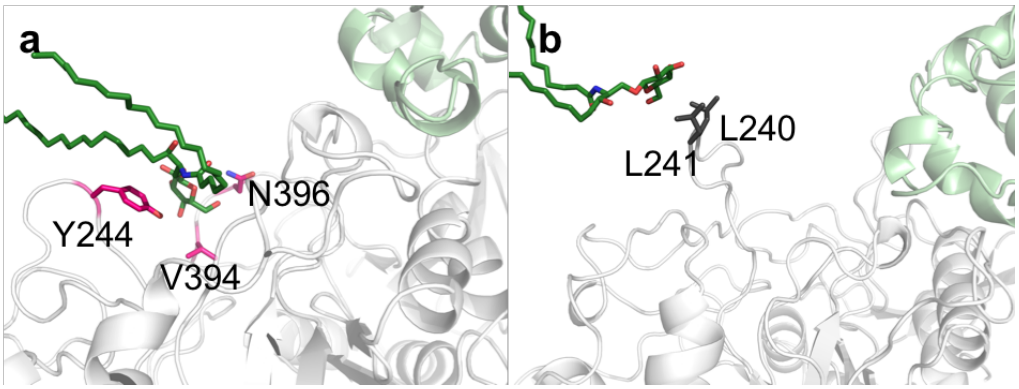

Figure 3.32: *Residues interacting with GluCer in simulation (a) 5a (CPX-Hel(N370S)) and (b) 5b (CPX-Ext(N370S)) at 500 ns. GluCer has been represented in green sticks and GCase in white with interacting residues in pink (5a) and grey (5b). Sap-c has been coloured in pale green. At 500 ns, only few interactions are observed in the binding site of the mutants.*

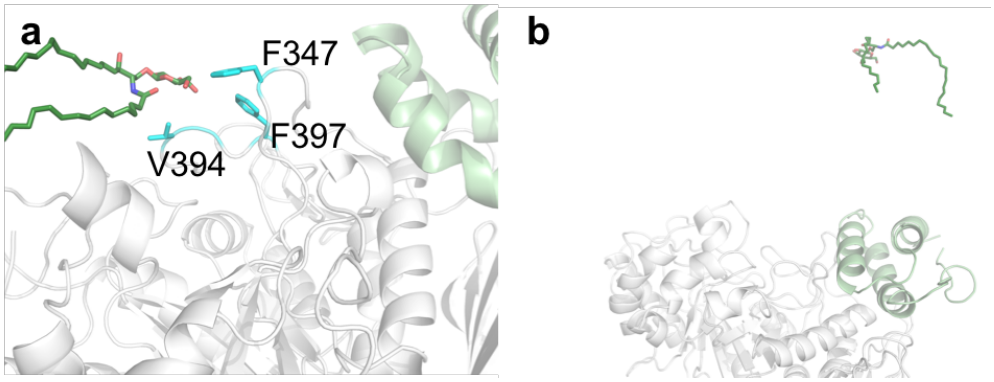

Figure 3.33: *Residues interacting with GluCer in simulation (a) 6a (CPX-Hel(L444P)) and (b) 6b (CPX-Ext(L444P)) at 500 ns. GluCer has been represented in green sticks and GCase in white with interacting residues in cyan (6a). Sap-C has been coloured in pale green. In simulation 6b GluCer was completely outside of the binding pocket at 500 ns. At 500 ns, only few interactions are observed in the binding site of the mutant in Simulation 5b, whereas in Simulation 6a GluCer has abandoned the binding site at this time.*

3.3.2.2.5. Protein- protein interactions

The protein-protein interactions are also affected in mutants. Many of the interactions produced in the wild types are found disrupted in the mutants. In simulations 5a (CPX-Hel(N370S)) and 5b (CPX-Ext(N370S)), where residue N370 has been mutated to Serine, the differences are notable (Fig. 3.34). In simulation 3a (CPX-Hel), the interaction between the residue H365 in helix 7 and residue D29 of Sap-C is consistent and stable during the simulation time, whereas it is absent in the mutant. In simulation 5a (CPX-Hel(N370S)), the interaction between residue D315 of GCase and K33 of Sap-C starts from 400 ns and partially recovers Loop-1 helicity. Interactions between residue K25 of Sap-C and residue N442 and D443 in the proximities of L444 are disrupted. However, the interactions between K25 and L444 and D445 are maintained. Some other protein-protein interactions are also broken, such as that between residue W348 and S43 (Sap-C) or between residues Q440 and E63 (Sap-C) (Fig. 3.36).

In simulation 5b (CPX-Ext(N370S), the docking between the proteins is poor. After approximately 400 ns of the simulation, Sap-C detaches almost completely from the mutated GCase. At this point, Sap-C is positioned near a completely deformed Loop-1 between residue K321 near Loop-1 and residues D29 and E26. Towards the end of the simulation, new interactions are formed between the mutant GCase and Sap-C, however not involving helix 7 (containing residue 370), for example, new interactions are formed between residue D51 of Sap-C and R44, Y487 or S465 of mutant GCase (N370S). Loop-2 is completely free. Interactions between K25 and L444 and surrounding residues are completely disrupted in this mutant simulation. Additional interactions are made between D29 of Sap-C and Y373 in the proximities of N370. This interaction is observed and is very stable in the corresponding wild type (Fig. 3.37).

In simulation 6a (CPX-Hel(L444P)) and 6b (CPX-Ext(L444P)), where L444 has been mutated to Proline the differences are also pronounced (Fig. 4.35). In simulation 6a (CPX-Hel(L444P)), interactions between residues K25 and P444 and D445 are disrupted from 600 ns onwards, although interaction with residue D443 is maintained from time 500 ns onwards. Interactions of Sap-C with Loop-1 of GCase are almost non-existent towards the

end of the simulation. Some interactions between Sap-C and Domain I and II of GCase are stable from 500 ns. For example, the interactions occurring between residue D51 of Sap-C and S12 and R44 of GCase (Fig. 3.38).

Finally, in simulation 6b (CPX-Ext(L444P)), interactions between residue K25 and residues P444 and other surrounding residues as D445 and D443 are completely lost. The disruption of those interactions makes Sap-C partially detached and move towards the upper part of helix 7 near Domain I. Interactions with Loop 1 are again almost non-existent. Some stable interactions are those between the residues S43 (Sap-C) and Q350 (GCase) and between residues D51 (Sap-C) and R353 or W357 (backbone) of GCase (Fig. 3.39).

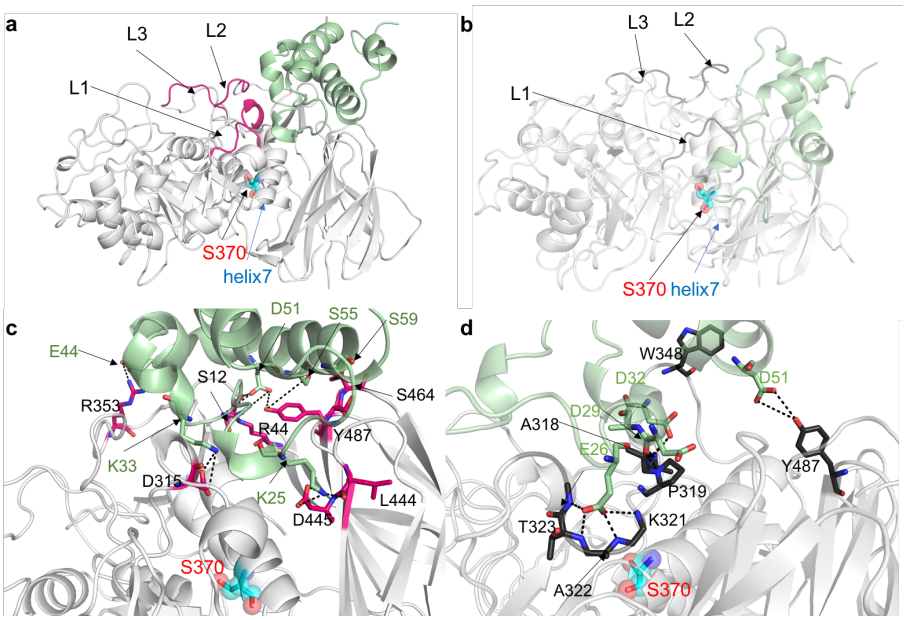

Figure 3.34*: Protein- protein interactions in simulation (a and c) 5a (CPX-Hel(N370S)) and (b and d) 5b (CPX-Ext(N370S)) are presented. Snapshot taken at 1000 ns of the simulation time. (a) General view of the complex GCase (white) and Sap-C (green), loops at the entrance of the binding site have been highlighted in pink (5a) and grey (5b) and been correspondingly labelled, mutated residue N370S has been coloured in cyan. (c and d) Detailed view of the protein binding site in simulation (c) 5a (CPX-Hel(N370S)) and (d) 5b (CPX-Ext(N370S)). GCase has been depicted in white and Sap-C in green. Interacting residues of GCase have been coloured in pink in 5a and dark grey in 5b. Mutated residue N370S has been coloured in cyan. In Simulation 5a some of the interactions remain the same as in the wild type whereas in Simulation 5b most of the analogous interaction observed in the binding site of the wild type are disrupted.*

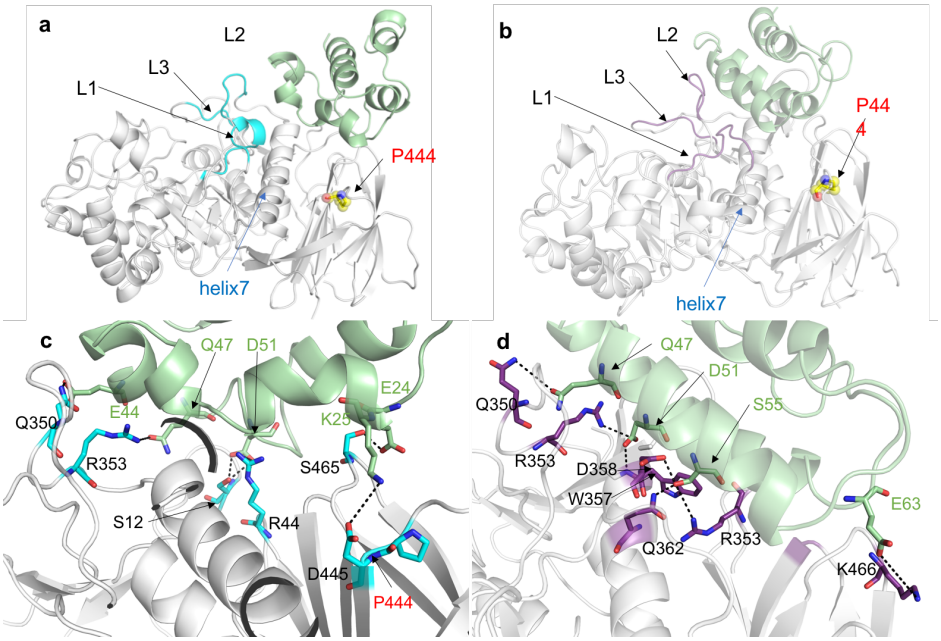

Figure 3.35: *Representation of the protein- protein interactions in simulation (a and c) 6a (CPX-Hel(L444P)) and (b and d) 6b (CPX-Ext(L444P)) . Snapshot taken at 1000 ns of the simulation time. (a) General view of the complex GCase (white) and Sap-C (green), loops at the entrance of the binding site have been labelled and highlighted in cyan (6a) and purple (6b). (c and d) Detailed view of the protein binding site in simulation (c) 6a (CPX-Hel(L444P)) and (d) 6b (CPX-Ext(L444P)). GCase has been depicted in white and Sap-C in green. Interacting residues of GCase have been coloured in cyan in simulation 6a and purple in simulation 6b. In simulation 6a and 6b many interactions occurring in the wild type are disrupted. This effect is pronounced in simulation 6b where all the interactions with Loop-1 have been lost.*

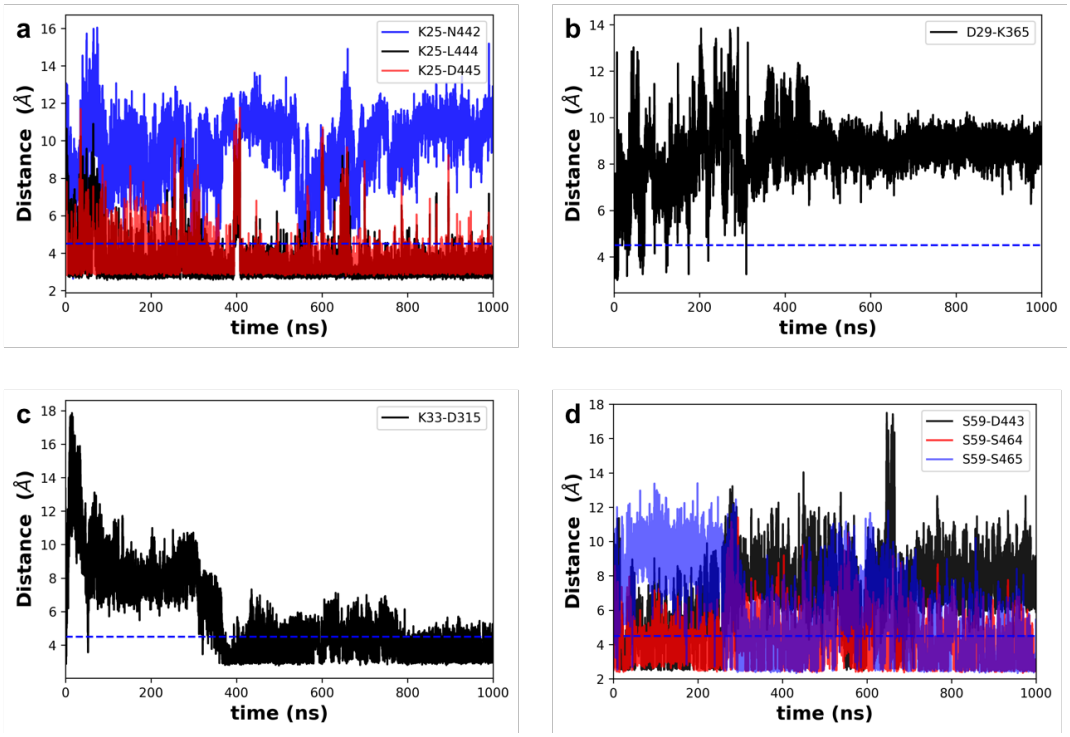

Figure 3.36: *Some of the protein- protein interactions measured along the time in simulation 5a.*

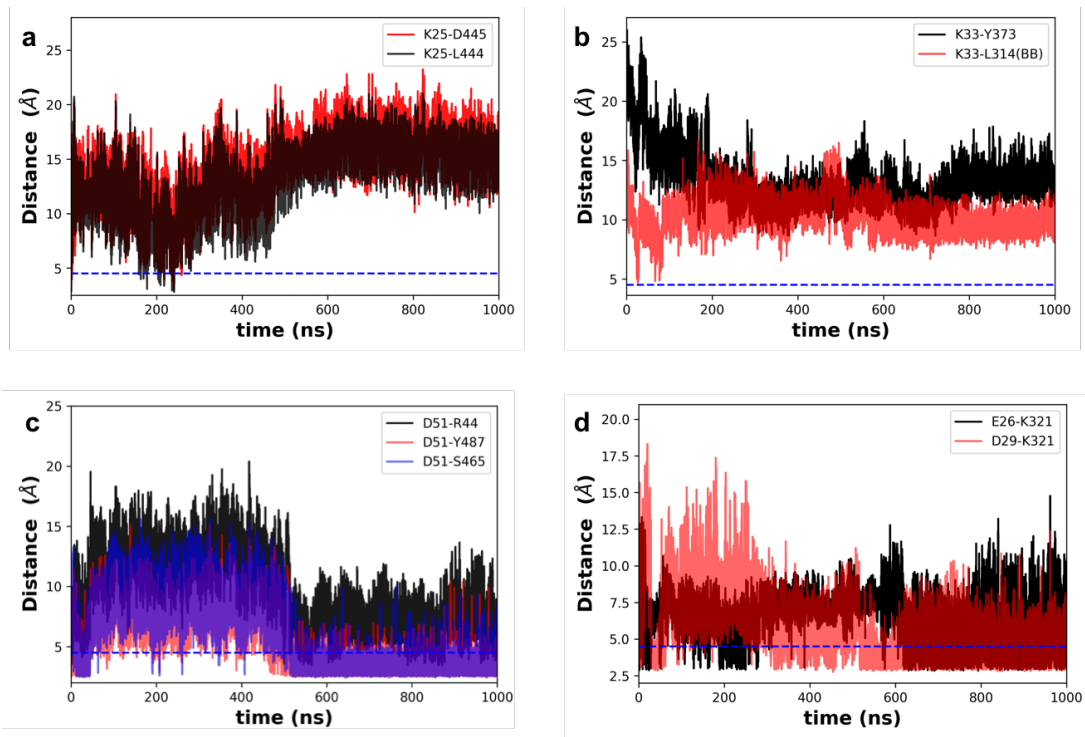

Figure 3.37: *Some of the Protein- protein interactions measured along the time in simulation 5b.*

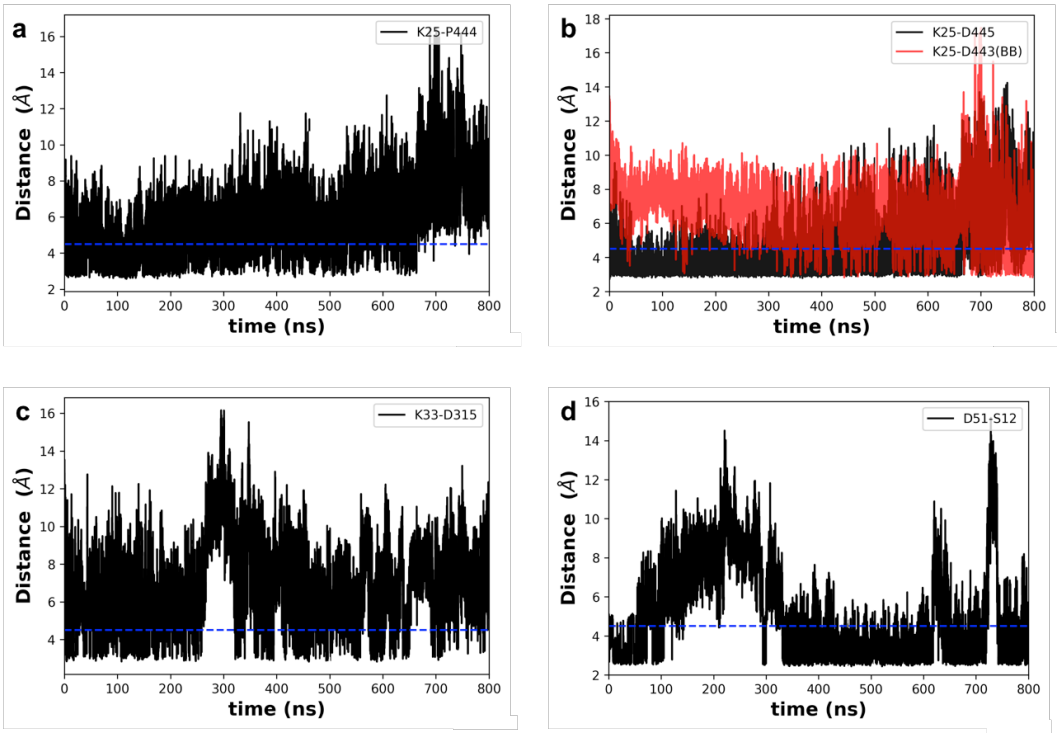

Figure 3.38: *Some of the protein- protein interactions measured along the time in simulation 6a.*

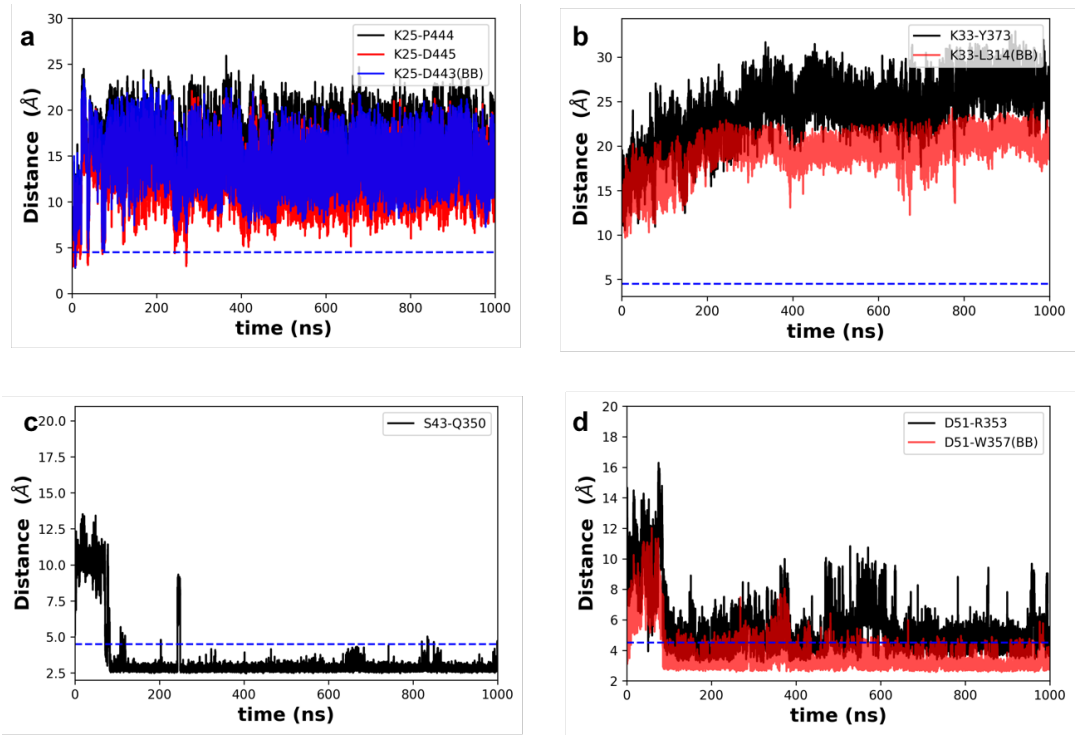

Figure 3.39: *Some of the protein- protein interactions measured along the time in simulation 6b.*

Sap-C	**5a-CPX-Hel N370S**	**6a-CPX-Hel L444P**	**5b-CPX-Ext N370S**	**6b-CPX-Ext L444P**
E24	-	S465	-	
K25	D445 L444(bb)	D445 P444(bb) (until 650 ns)	-	-
E26	-	-	K321 (from 650 ns)	-
D29	-	-	K321	-
D32	N362 (until 300 ns)	N362 (until 600 ns)	L317 (bb)	-
K33	D315 (from 350 ns)	D315 (until 250 ns)	-	-
S36	-	-	-	-
S43	-	-	-	Q350
E44	R353	Q350	-	-
Q47	-	R353 (from 400 ns)	-	Q350 N353
D51	Y487 R44 S12 (from 350 ns)	R44 (from 300 ns) S12 (from 350 ns)	S465 (from 500 ns) Y487 (from 500 ns)	R353 W357 (bb)
S55	Y487	-	-	D358 Q362 (after 750 ns) R44 (after 750 ns)
S56	-	-	K466 (from 500 ns)	-
S59	S464 S465	-	-	-
L62(bb)	-	-	-	-
E63	K441	-	-	K466

Table 3.6: *Summary of protein-protein interaction in four simulations 5a, 5b, 6a and 6b. The abbreviation bb stands for 'backbone'.*

3.3.2.2.6. A comparison of wild-type N370 and L444 with mutants N370S and L444P

N370 (simulation 3a, complex in active conformation)

At the start of the simulation, N370 interacts with the side chains of residues T369 and S366. After 25 ns of simulations, the side chain of N370 located in helix-7 of Domain III flips towards β-strand 7 forming stable hydrogen bonds with the backbone atoms of residues W378 and G377, which are maintained throughout the course of the simulation.

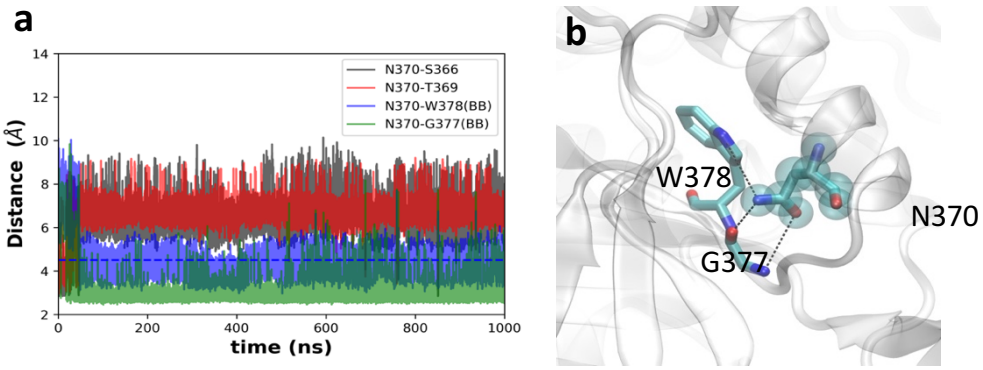

Figure 3.40: *(a) Distance between residue N370 and residues S366, T369, W378 and G377, in simulation 3a (CPX-Hel). (b) Snapshot of the interactions between N370 and W378 (bb) and G377 (bb) in simulation 3a (CPX-Hel) at 1000 ns.*

L444 (simulation 3a, complex in active conformation)

Backbone of residue L444 forms stable hydrogen bonds with the side chain of the residue K25 of Sap-C and N442, as well as with the backbone of residue N442. The side chain of the residue L444 lies in a hydrophobic pocket formed between the two β-sheets of Domain II. Residues A446, V460, V468, L470, I60, L65 also form a part of this pocket.

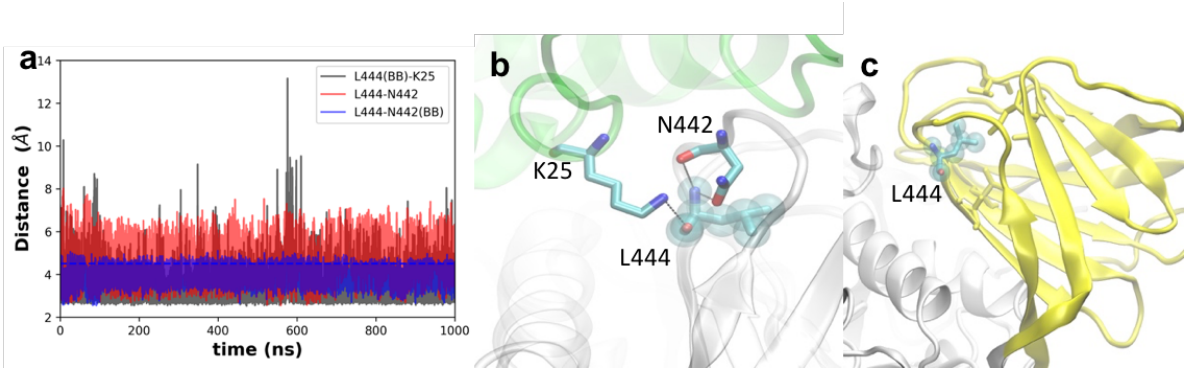

Figure 3.41: *(a) Distance between residue L444 (bb) and residues K25 of Sap-C, N442 and N442 (bb), in simulation 3a (CPX-Hel). (b) Snapshot of the interaction between residues L444 (bb) and K25 of Sap-C and N442 in simulation 3a (CPX-Hel) at 1000 ns. (c) Side chain of residue L444 lies in a hydrophobic pocket between the two Beta sheets of Domain II.*

N370 (simulation 3b, complex in inactive conformation)

In the inactive conformation, N370 interacts with residue S366 and T369 in the same helix and forms very stable hydrogen bonds (Fig 3.42). N370 also interacts with the side chain of residue W312 in Loop-1. It should be worth mentioning here that difference in the conformation of Loop-1 is one of the markers that differentiate the active and inactive states of GCase. Thus an interaction between the side chain of N370 and residues in Loop-1 is a direct link that plays a role in activation.

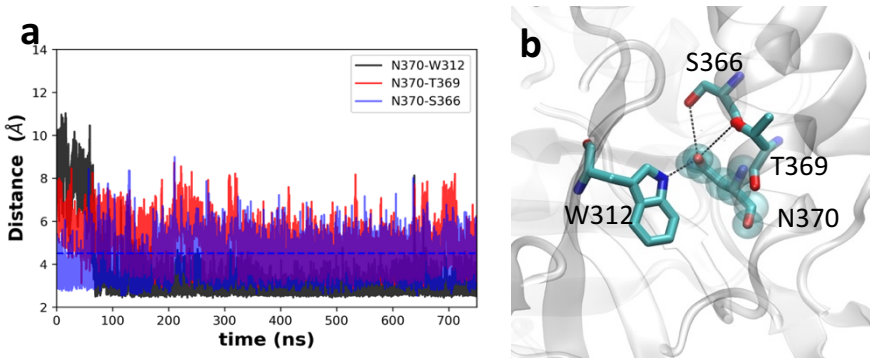

Figure 3.42: *(a) Distance between residue N370 and W312, S366 and T369 in simulation 3b (CPX-Ext). (b) Snapshot of the interaction between N370 and W312, S366 and T369 in simulation 3b (CPX-Ext) at 1000 ns.*

L444 (simulation 3b, complex in inactive conformation)

Backbone of residue L444 forms stable hydrogen bonds with the side chain of the residue K25 of Sap-C and with the backbone of residue N442 and K441. The side chain of residue L444 lies in a hydrophobic pocket formed between two β sheets of Domain II. Residues A446, V460, V468, L470, I60, L65 also contribute in part to this pocket.

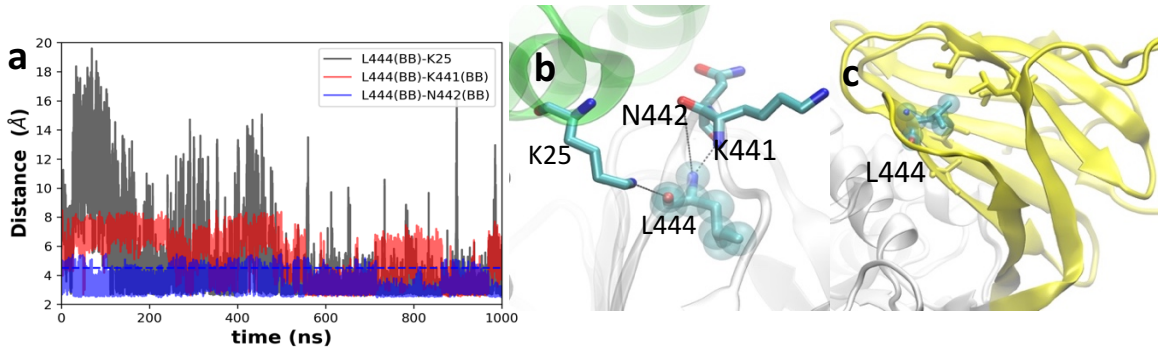

Figure 3.43: *(a) Distance between residue L444 (bb) and the side chain of K25 of Sap-C, K441 (bb) and N442 (bb), in simulation 3b (CPX-Ext). (b) Snapshot of the interaction between residues L444 (bb) and K25 of Sap-C, K441 (bb) and N442 (bb) in simulation 3b (CPX-Ext) at 1000 ns. (c) Side chain of residue L444 lies in a hydrophobic pocket between two β sheets of Domain II.*

Mutant S370 (simulation 5a, N370S complex in active conformation)

The mutant S370 forms a stable hydrogen bond with the side chain of residue S366 in the same helix. It also interacts with the backbone of residue V375 in β strand 7 up to 550 ns. Towards the end of the simulation, the mutated residue S370 interacts with residue W312 in Loop-1.

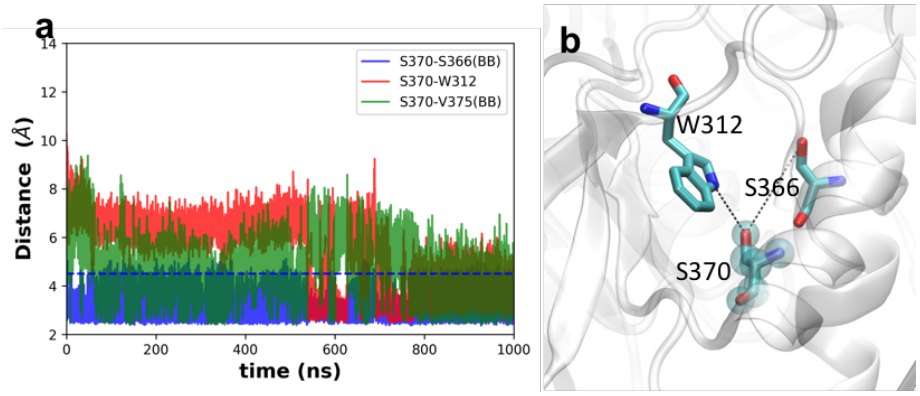

Figure 3.44: *(a) Distance between residue S370 and residues W312, S366 and V375 (bb) in simulation 5a (CPX-Hel(N370S)). (b) Snapshot of the interaction between N370 and W312 and S366 in simulation 5a (CPX-Hel(N370S)) at 1000 ns.*

L444 (simulation 5a, N370S complex in active conformation)

Backbone of the L444, forms stable hydrogen bonds with the sidechain of the residue K25 of Sap-C and with the sidechain of residue D443 and the backbone of residue N442. The sidechain of residue L444 is positioned in a hydrophobic pocket formed between the two $\beta-$sheets of Domain II. Residues A446, V460, V468, L470, I60, L65 are part of this pocket.

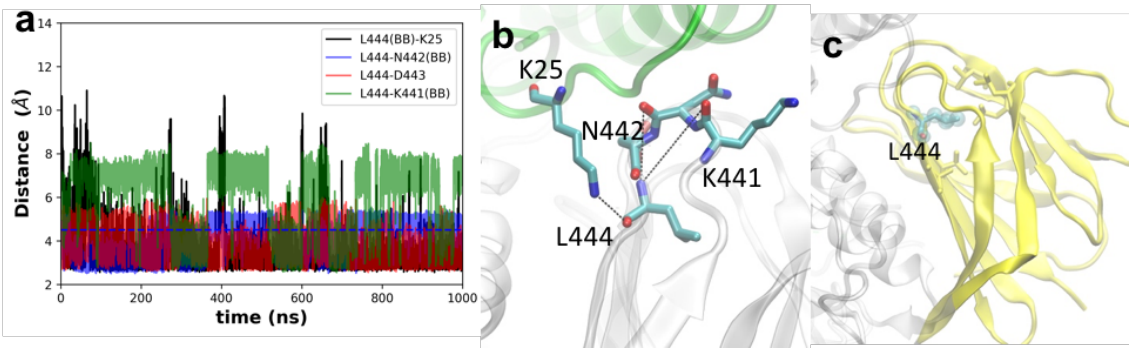

Figure 3.45: *(a) Distance between residue L444 (bb) and residues K25 of Sap-C, K441 (bb), N442 (bb) and D443, in simulation 5a (CPX-Hel(N370S)). (b) Snapshot of the interaction between residues L444 (bb) and K25 of Sap-C K441 (bb) and N442(bb) in simulation 5a (CPX-Hel(N370S))at 1000 ns. (c) Sidechain of residue L444 lies in a hydrophobic pocket between the two β sheets of Domain II.*

Mutant S370 (simulation 5b, N370S complex in inactive conformation)

As observed in the active conformation, the mutant S370 forms a stable hydrogen bond with the side chain of the residue S366 in the same helix, but also with the residue T369 during the entire simulation. S370 also interacts with W378 in β strand 7. From ~ 500 ns the mutated residue interacts with R285 in helix 5, and forms a stable interaction until the end of the simulation.

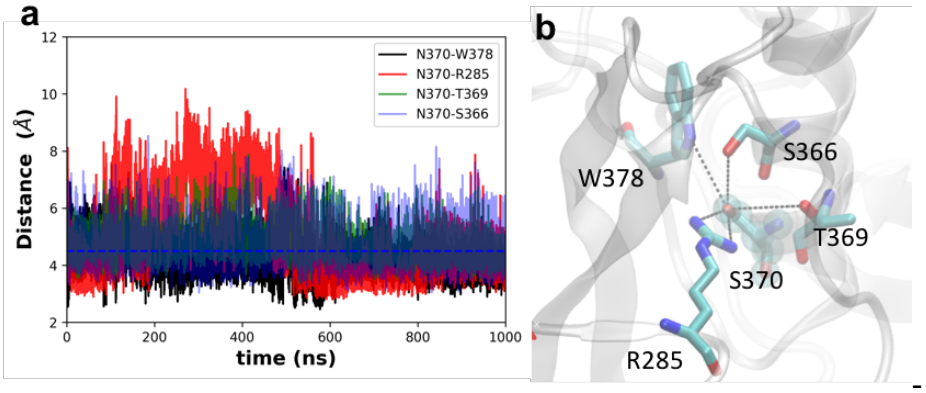

Figure 3.46: *(a) Distance between residue S370 and residues W378, S366, T369 and R285 in simulation 5b (CPX-Ext(N370S)). (b) Snapshot of the interaction between N370 and W378, S366, T369 and R285 in simulation 5b (CPX-Ext(N370S)) at 1000 ns.*

L444 (simulation 5b, N370S complex in inactive conformation)

Backbone of residue L444 does not form a hydrogen bond with residue K25 of Sap-C. The interaction is disrupted along with the other interactions that K25 of Sap-C makes with surrounding residues. The backbone atoms of L444 form a stable bond with the sidechain of residue D443 and the backbone of residue N442. As observed in the wild type, the sidechain of residue L444 is positioned in a hydrophobic pocket formed between the two β-sheets of Domain II. Residues A446, V460, V468, L470, I60, L65 form also part of this pocket.

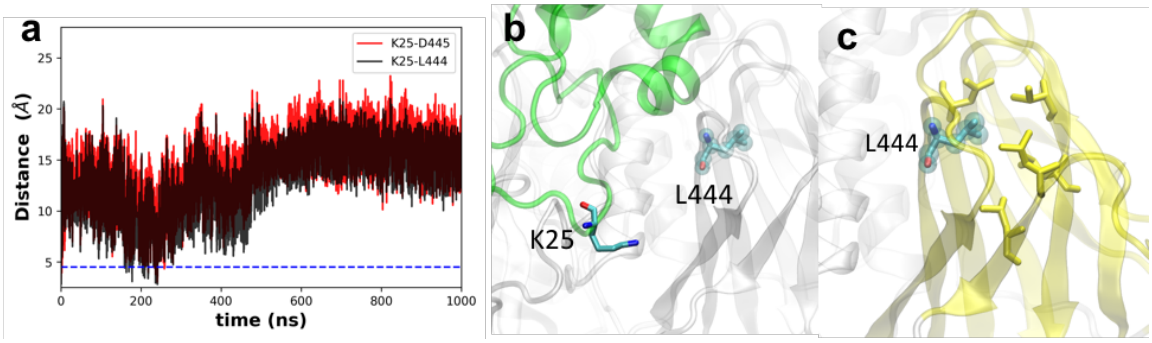

Figure 3.47: *(a) Distance between residue L444 (bb) and residues K25 of Sap-C, and D445 and K25 of Sap-C in simulation 5b (CPX-Ext(N370S)). (b) Snapshot of the interaction between residues L444 (bb) and K25 of Sap-C, K441 (bb) and N442 (bb) in simulation 5b (CPX-Ext(N370S)) at 1000 ns. (c) Sidechain of residue L444 lies in a hydrophobic pocket between the two β sheets of Domain II.*

N370 (simulation 6a, L444P complex in active conformation)

N370 interacts with the backbone of the residues G377 and V375 in the β strand 7 via hydrogen bonds. At 500 ns the side chain of residue N370 flips and establishes hydrogen bonds with residues W312, S366 and W378. However, this is transitional and only lasts for about 100 ns.

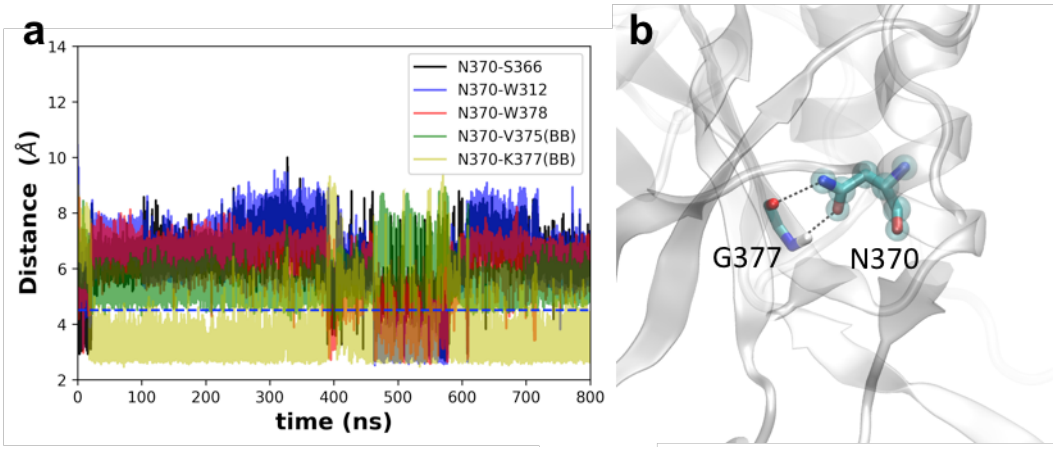

Figure 3.48: *(a) Distance between residue N370 and residues W378, S366, W312, V375 and G377 along simulation 6a (CPX-Hel(L444P)). (b) Snapshot of the interaction between N370 and G377 in simulation 6a (CPX-Hel(L444P)) at 1000 ns.*

Mutant P444 (simulation 6a, L444P complex in inactive conformation)

The interaction between P444 and residue K25 of Sap-C, which is present in rest of the simulations, is disrupted in simulation 6a from ~ 600 ns of the simulation. The backbone of P444 forms hydrogen bond interactions with residue N442 and with residue D445. P444 lies inside a hydrophobic cluster in the middle of the two beta sheets that forms Domain II. These hydrophobic interactions are maintained in the L444P mutant.

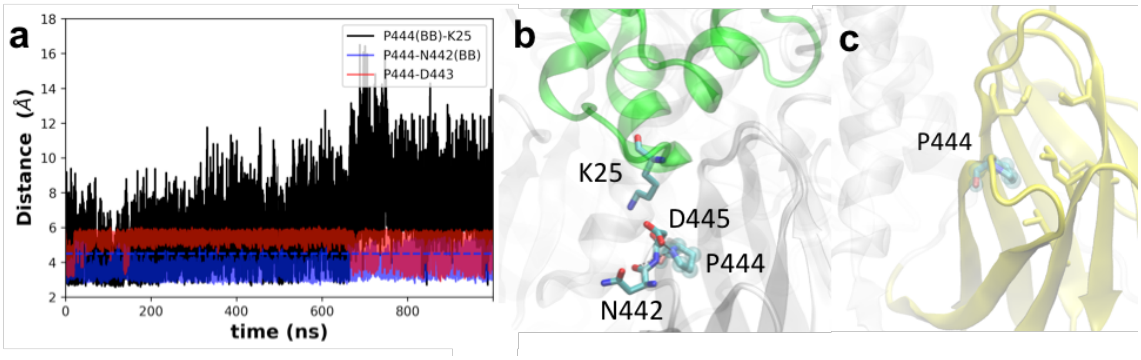

Figure 3.49: *(a) Distance between residue P444 (bb) and residues K25 of Sap-C, D443 and N442 in simulation 6a (CPX-Hel(L444P)). (b) Snapshot of the interaction between residues P444 (bb) and K25 of Sap-C, D443 and N442 (bb) in simulation 6a (CPX-Hel(L444P)) at 1000 ns. (c) Sidechain of residue L444 lies in a hydrophobic pocket between the two β sheets of Domain II.*

N370 (simulation 6b, L444P complex in inactive conformation)

N370 interacts with residue S366 and T369 in the same helix along the simulation time in a stable hydrogen bond. N370 forms a hydrogen bond with residue R285 until 400 ns after which it gets broken. From 400 ns onwards, two other interaction are formed, between residues W312 in Loop-1 and H374 in the loop connecting helix 7 and β strand 7.

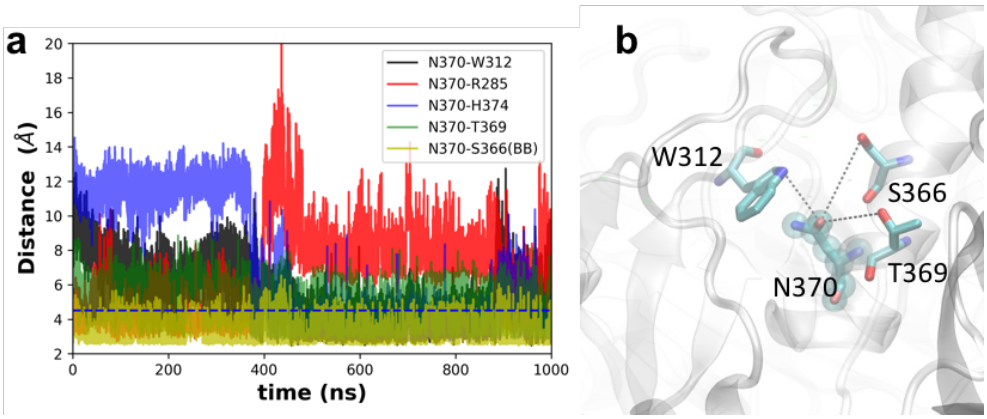

Figure 3.50: *(a) Distance between residue N370 and residues R285, W312, S366, T369 and H374 along simulation 6b (CPX-Ext(L444P)). (b) Snapshot of the interaction between N370 and W312, S366 and T369 in simulation 6b (CPX-Ext (L444P)) at 1000 ns.*

Mutant P444 (simulation 6b, L444P complex in inactive conformation)

The interaction between P444 and K25 of Sap-C, which is present in rest of the simulations, is disrupted in simulation 6b. The backbone of P444 forms hydrogen bond interactions with residue N442 in the first half of the simulation, after which it begins to interact with residue K441 in a less stable interaction.

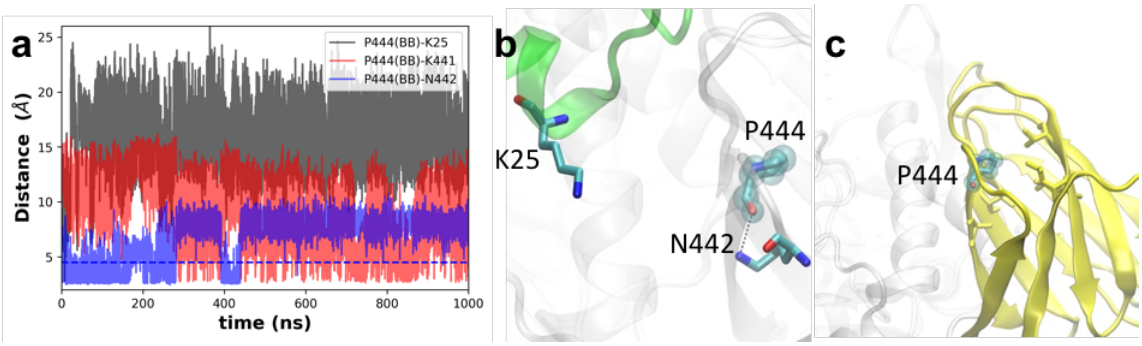

Figure 3.51: *(a) Distance between P444 (bb) and residues K25 of Sap-C and N442 (bb) in simulation 6b (CPX-Ext(L444P)). (b) Snapshot of the interaction between residues P444 (bb) and N442 (bb) in simulation 6b (CPX-Ext(L444P)) at 750 ns, (interaction with residue K25 of Sap-C does not occur in this simulation). (c) Mutation to residue P444 disrupts the hydrophobic pocket between the two β sheets of Domain II.*

3.4. **Discussion**

Our motivation in the present research has been to understand the dynamics of GCase at atomistic level and the role of its different components, in order to give a plausible structural explanation of the implications of the different mutations found in Gaucher's Disease. Molecular Dynamics allows us to follow the dynamic evolution of a biological system. Simulations of >500 ns can highlight the dynamics of a membrane protein. Simplified techniques such as Coarse-Grained help to accelerate the process.

In spite of many in vitro experiments, there is no evidence of an interaction between GCase and Sap-C in solution, i.e. in absence of a lipid bilayer.[123,124,125] However, the interaction of both is recovered when lipids are added. Negatively-charged lipids are abundant component of the intralysosomal membrane[126] and are required for the activation of GCase by Sap-C.[123,124,125] In 2007, Jean-René Alattia et al. studied activation by Sap-C and observed that GCase hydrolyses its substrate at the bilayer level with the help of SapC within a complex at the membrane surface.[127] They proposed a "liftase" mode of action for Sap-C according to which GCase would not be able to penetrate the membrane, thus GluCer ought to be ''lifted'' for proper docking to the active site. Using a membrane-bound fluorogenic substrate analogue, they observed an increase in the GCase of 17-fold in presence of Sap-C. The group postulated that this SapC-induced enhancement in the enzymatic activity should also be related to a greater intrinsic activity of GCase within an activator complex, probably involving a conformational change in the hydrolase, as observed in the case of pancreatic lipase.[128]

Alattia et al. finding highlighted the importance of studying the activation process within a lipid membrane. Based on that hypothesis, we conducted four MD runs where we have simulated GCase alone, with its substrate, as an entire complex with its facilitator protein Sap-C and the facilitator protein alone, always in presence of a lipid membrane. Coarse-grained simulations gave us the opportunity to observe the lipid self-assembly process. Quality controls demonstrated the correct formation of the membrane. GCase and Sap-C anchored to the membrane as peripheral membrane protein in all the simulations, as it was

expected.

The CG coordinates of all the simulations were transformed to atomistic. Additional 1000 ns atomistic simulations were conducted in order to understand the conformational changes in more detail. The atomistic simulations included GCase in a membrane environment, GCase along with it substrate in active and inactive conformation, the entire complex with Sap-C and substrate in both conformations and the two most clinically important mutants in complex with Sap-C, using both conformations of the enzyme, active and inactive.

The structural stability of the systems was assessed by measuring the RMSD values. The results of RMSD were expected and followed a trend. The values of RMSD were more stable when GCase was simulated along with Sap-C (Simulation 3a and 3b). The mutated proteins exhibited a similar trend when in complex with Sap-C and when GCase is simulated alone. These results mean that Sap-C was able to stabilise the enzyme during the simulation of the complex. The complexes of the two mutants, in both conformations were more unstable than their wild type counter part. When GCase is simulated in extended (inactive) conformation, it exhibits higher RMSD values than when it is simulated in active conformation, indicating a higher conformational flexibility.

The RMSF values were analysed in the context of loop dynamics, interactions within the binding site and protein- protein interaction. The highest peaks of RMSF corresponded to surface loops whereas the core structure was stable. Some differences were observed in the loops present at the entrance of the binding site. Simulations of GCase alone (2a and 2b) displayed higher RMSF in Loop-1. Subsequent analysis showed that Loop-1 partially lost its helical form during Simulation 2a (GCase-Hel and GluCer), whereas it extended towards helix 7 in simulation 2b (GCase-Ext and GluCer). In simulation 3a (CPX-Hel), Loop-1 conserved its helicity during the entire simulation due to the restraint placed by interaction with residue K33 of Sap-C. Finally, in simulation 3b (CPX-Ext), Loop-1 does not change its extended conformation. Loop-2 displayed RMSF values above 2 Å in two simulations, 2a (GCase-Hel and GluCer) and 3b (CPX-Ext). In simulation 2a (GCase-Hel and GluCer), the loop goes from an active conformation to become embedded inside the lipid membrane, and 3b (CPX-Ext), where the loop gets tucked in a hydrophobic pocket

under Sap-C. Only during simulation 3b, Loop-3 displayed high RMSF. During this simulation the loop goes from a closed conformation, where the side chain of residue R395 points towards the inside of the binding pocket, to an open conformation. In both simulations of the wild type active conformation (2a and 3a), the RMSF value of Loop-4 peaks above 2 Å. Analysing the interactions occurring within the binding site, this loop was observed to interact with the substrate within the active site in the active conformation.

In the simulations of mutants in extended conformation (5b and 6b), Loop-1 displays high RMSF values around residues near helix 6. Helix 6 loses its helical conformation during the simulation and extends towards helix 7 and the membrane. A similar behaviour of this loop is also observed in the mutant active conformation 6a (CPX-Hel(L444P)) and partial (only the lower part of the helix) unfolding in the wild type inactive conformation when run without Sap-C, in simulation 2b (GCase-Ext and GluCer). The deformation of the helix 6 and part of the Loop-1 result in structural instabilities at the protein- protein interface, which partially detaches Sap-C from GCase.

Analysis of the dynamics of the loops at the entrance of the binding site in GCase shed light on the influence of the facilitator protein Sap-C. We have focused our efforts in analysing Loop-1, 2 and 3, as these are the activation loops, whereas Loop 4 and 5 have a structural and role in molecular recognition. Simulations of active mutants, although they do not exist in nature have helped us to understand part of the activation/ inactivation process.

Loop-1 (H311-P319) has been reported to adopt a helical conformation in the active state of the enzyme, as demonstrated in different crystallization experiments[27]. In the helical conformation, the side chain of residue D315 points towards the outside of the binding site. The helicity of Loop-1 is partly lost when GCase is simulated without Sap-C, in simulation 2a (GCase-Hel and GluCer) (Figure 3.13). In simulation 3a (CPX-Hel), the helicity is maintained because of a stable interaction, that lasts the entire simulation, between the residues D315 of GCase and K33 of Sap-C. In simulation of the mutant N370S active conformation 5a, the helical conformation of the loop is completely lost at the beginning of the simulation. This recovers partly once a hydrogen bond interaction is formed between

D315 and K33. This is also observed in simulation 6a (CPX-Hel(L444P)). At the beginning of the simulation, Loop-1 loses it helical form. This is recovered after the formation of a hydrogen bond between D315 and K33 of Sap-C. Following this bond formation, the helicity is maintained due to additional interactions with H365 an S366 in helix 7. In simulation 2a, where active GCase is simulated without Sap-C, Loop-1 partially loses helical form and establishes interactions with some residues of Loop-2, namely W348 and K346. In simulation 3b (CPX-Ext), residue K33 of Sap-C interacts via hydrogen bonds with the backbone atoms of L314 and Y373 throughout the simulation. It is important to note that these two residues surround the residue D315. In this thesis, we propose a helication mechanism of Loop-1 based in the interaction of residue D315 of GCase and K33 of Sap-C. In our model of activation residue K33 of Sap-C would form an ion pair that would take D315 from interacting with residues of Loop-2 (extended conformation) to interact with residues in helix 7.

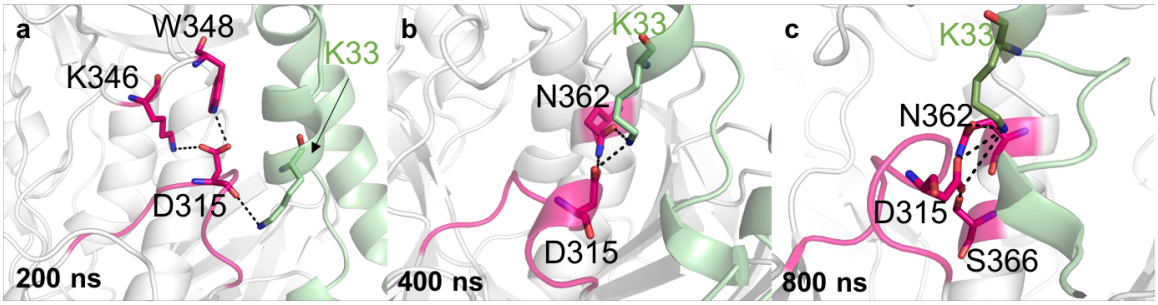

Figure 3.52: *Evolution of Loop-1 in simulation 6a (CPX-Hel(L444P)) at (a) 200, (b) 400 and (c) 800 ns. GCase has been depicted in white with the interacting residues in pink and Sap-C has been coloured in green. At the beginning of the simulation, D315 is extended towards Loop-2 (characteristic of Loop-1 of the inactive conformation). As the residue starts interacting with residue K33 of Sap-C, Loop-1 orients towards helix-7; a characteristic of the active conformation.*

An experimental study carried out by Joel L. Sussman's group in 2005 detected significant structural changes in both Loop-2 and Loop-3 upon the binding of the irreversible inhibitor CBE (Conduritol-B-Epoxide).[129] The said study suggests that both loops act as a lid of the active site, thus highlighting the importance of both in the activation of the enzyme. Both Loops presented a closed conformation upon CBE binding. In Loop-2 there were no

prominent structural changes in the structure of the loop, although in the inactive conformation the loop seemed closer towards the binding site. In Loop-3 the structural changes were greater with bulky residues R395 and F397 pointing towards the binding site. Analogous loops in the enzyme glycosyl transferase have also been reported to be important for its activation.[130]

In this thesis, we have described above how Loop-2 adopts an open conformation in presence of Sap-C. In the wild type simulations 3a (CPX-Hel) and 3b (CPX-Ext), residue W348 is tucked in a hydrophobic pocket made by Sap-C at the interface of both proteins. In the mutants except in 5a (CPX-Hel(N370S)), W348 is not tucked inside this binding pocket. Among all the simulations of extended conformation only 3b (CPX-Ext) presented an open conformation of Loop-3. In the rest of the simulations of the extended conformation, namely 2b (GCase-Ext and GluCer), 5b (CPX-Ext(N370S)) and 6b (CPX-Ext(L444P)), residue R395 in Loop-3 formed a hydrogen bond interaction with residue E340 that completely obstructed the binding pocket. In this thesis, we would like to propose an opening mechanism for both Loop-2 and Loop-3 based on these observations. The opening mechanism of Loop-2 would depend on Sap-C, and relies on the fact that residue W348 of Loop-2 gets trapped in a hydrophobic pocket upon the binding of the facilitator protein. Tethering of residue W348 by Sap-C would not only produce the opening of Loop-2, but also disrupt some important interactions between Loop-2 and Loop-3. This would also prompt the opening of Loop-3. The change in conformation of Loop-3 coincides with the stabilization of the spatial position of W348. This has been illustrated in Figure 3.53.

The structural differences found at the protein- protein interface and the stability of interactions in different simulations, reflect the dynamics of the protein- protein recognition. Proteins do not fit in a static manner as building blocks but via a flexible and evolving process. Nonetheless, the disruption of some of these interactions can alter the process and makes protein- protein interaction impossible.

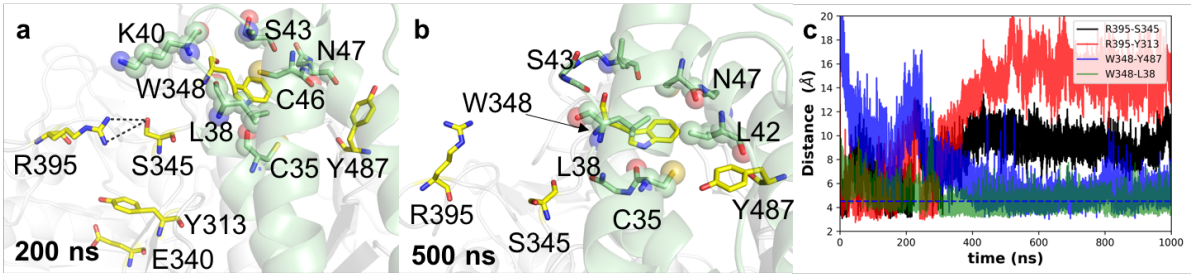

Figure 3.53: *Stabilization of the side chain of W348 inside the hydrophobic pocket formed by Sap-C. This coincides with the opening of Loop-3 in simulation 3b.* ***(a)*** *Interactions of residues W348 and R395 at 200 ns, (**b**) interactions of residues W348 and R395 at 500 n., GCase has been depicted in white with interacting residues in yellow and Sap-C in green.* ***(c)*** *Distance between pair of residues (R395-S345, R395-Y313, W348-Y487, W348-L38) in simulation 3b.*

Mutant L444P was reported to present a dramatic decrease in the activation by Sap-C, suggesting that the mutation can directly hamper the formation of the activator complex.[131] In this thesis, present a structural explanation for this reduction in the activation of the mutant. The mutation of L444 to Proline, a less bulky and more rigid residue, possibly prevents the interaction with residue K25 to occur. As a result, there are no interaction between Sap-C and GCase.

Mutant N370S has been reported to have a poor association with Sap-C unless the membrane is highly enriched with anionic phospholipids (>50%),[42] whereas in the wild type an extensive enzyme- Sap-C association occur when the anionic phospholipid threshold exceed 10%.[132] We observed that mutant N370S appeared to destabilize GCase, where Loop-1 extends towards helix 7, hampering the association with Sap-C.

On the other hand, some residues of Sap-C seemed to be key in making protein- protein interaction. For example D29 and D32 make stable interactions with the hydrogen bond donor residues in helix 7: N362, H365, T369 and Y373. K33 and K25 also seem to be decisive for the protein-protein interaction. K25 interacts with L444 and surrounding residues whereas K33 interacts with D315. Furthermore, serines and tyrosines at the edge of Domain II, facing helix 7 also form substantial interactions at the interface.

Molecular dynamics simulations allow us to study structural dynamics and the activation mechanism of GCase. However to thoroughly explore the conformational space of the protein further analysis is required.

CHAPTER 4:

KINETIC STUDIES

CHAPTER 4: KINETIC STUDIES

4.1. Introduction

Proteins are not static, but dynamic entities that exist as collections of interconvertible conformations in thermal equilibrium.[133] Thermodynamic fluctuations allow the protein to visit other conformations in a multidimensional free energy landscape (FES).[134,135,136,137] Conformational FES of proteins are rough surfaces with multiple hills and valleys of different depths and heights. Each valley represents a low-energy state and is populated by an ensemble of related conformations called substates. A substate is formed by a set of energetically and structurally similar conformations that are called microstates.[133] Substates are separated by energy barriers of different heights. If few substates are separated by low energy barriers they can be considered a new substate, if they are collectively separated from other distant substates by higher barriers. Thus, Conformational FES has a multilevel organization.[138] When the protein transits form one substate to another there are some intermediate, unstable and scarcely populated conformations named metastable states. Metastable states are the equivalent to the transition state in a chemical reaction and would be impossible to isolate by physical techniques.[138]

The conformational changes that a protein can undergo vary from the vibration of a bond or the movement of a small group of atoms to concerted movements in which the entire protein is involved.[133] The former are fast motions and occur in the ps to ns timescale. Fast motions allow the protein to visit close conformations within its substate. The latter are slow motions that occur in the μs to ms timescale. Slow motions allow the protein to explore distant substates in the conformational FES.[134,133,138] Slow motions underlie the designated function of proteins.[139] Big efforts have been made to understand slow conformational motions to unravel some important events such as protein folding, enzyme activation or ligand binding.[133,134]

Atomistic MD is still not sufficient to draw a complete picture of the FES of a protein.[140] Since slow motions occur in the μs to ms timescale, it is difficult to obtain a trajectory that

surpass a few μs in a reasonable real time. In addition, the conformational motions in a MD simulation are highly dependent on the initial structure, which can bias the conformational search of the protein.[133] Simple visual inspection of the trajectory and traditional measures such as RMSD can mislead the interpretation of the conformational subspace.[140] Thus, new techniques of sampling and analysis of the macromolecular transitions, that supplement the MD data with statistical significance, should be introduced. The starting point of those methods should be a dimension reduction that simplify the MD output.[140,141]

Anharmonic Conformational Analysis (ANCA) performs higher order statistics of conformational motions associated with the data sampled during the MD simulations, to identify substates.[142,143] ANCA relies on the demonstrated fact that slow motions accountable for the protein function show high levels of anharmonicity.[144] By focusing on the anaharmonicity ANCA is able to identify those conformational fluctuations that allow proteins to explore distant substates. ANCA links identified substates and conformational transitions to biophysical relevant characteristics highlighting the importance of those in the designated protein function.[143]

Discrete state kinetics models such as Markov State Models (MSMs) have been shown to be successful to reconstruct protein conformational FES and transition pathways among substates, including the identification of transient states or metastable states.[140,145] MSMs are able to extract information from a set of simulations of different starting points and reconstruct protein slow motions, even though those simulations are much shorter than the process to study.[146] The resulting model is a network of discrete states (markovian) separated by probabilities of transitions. Also called transition network, MSMs produce efficient, easily readable models to study the conformational transition of proteins.[145,147]

In this thesis, we have combined ANCA and MSMs to analyse the MD data from the wild type simulations in order to extract kinetic relevant information from the activation process of the enzyme GCase. Both techniques will be explained into greater depth in following sections.

4.1.1. Anharmonic Conformational Analysis (ANCA)

Anharmonicity is demonstrated to be an essential characteristic of time-dependent conformational fluctuation of proteins.[144,148,149] Anharmonic events are long (μs to ms timescale) and rare events accountable for the designated function of biomolecules. ANCA uses fourth order statistics to explore anharmonicity of such events and thus characterise positional fluctuations responsible for conformational transition of proteins.[143,134] Internal motions are then summarised using a small number of dominant anharmonic modes. Conformational space is partitioned in a series of substates and conformational transition in a multilevel motion hierarchy.[138]

4.1.1.1. Kurtosis

Kurtosis is a fourth order statistic measure of the anharmonicity in atomic fluctuation. For a real random variable, Kurtosis (κ) is defined as:[133]

$$\kappa\,(q) = \frac{E\{(q-\mu)^4\}}{\sigma^4} \tag{4.1}$$

where q denotes a real random variable, μ and σ represent respectively the mean and standard deviation of q and E{q} the expected value of the variable q. For unimodal distributions, the kurtosis is a measure of the peak or the proportion of the weights in the tails. For a Gaussian distribution with zero mean the value of κ is equal to 3. A super-Gaussian, with heavier tails and more peaked would be greater than 3. Values of κ lower than 3 would define a sub-Gaussian distribution, a less peaked distribution. Either the Cartesian coordinates or dihedral angle selections can be used to calculate the κ in ANCA.[133]

To study the evolution of the κ throughout the simulation time and spot possible events, a sliding window analysis should be carried out. For that, we can use an exponential window located at time t, a weight:[150]

$$W_k = \alpha e^{-(t-k)/\tau} \tag{4.2}$$

where k is a frame from the past (k < t) and τ is the time constant for exponential weight decay. The weights are natural solution for smoothing any considerable fluctuations observed in time evolving properties (like κ) along the simulation time.[150]

Kurtosis of the positional deviations can be projected onto a system built for each Cα. Thus, various researches have demonstrated that non-Gaussian distributions (either sub- or super- Gaussian) are associated with functionally relevant regions of proteins.[142]

When the distributions of atom deviations are Gaussian-like, the ANCA basis vectors which maximize variance, align well with the intrinsic orientation of the data.[133] However, when the atomic distributions combine sub- and super- Gaussian distributions, the intrinsic orientations of the data could be non-orthogonal, being necessary higher-order correlations.[133,150] To overcome those problems, namely higher-order correlations and non-orthogonality, ANCA introduces four modules for the characterization of anharmonic modes of motion in the conformational landscape.

4.1.1.2. Solving Spatial and Temporal correlations.

ANCA introduces four core modules for the analysis of MD simulations. These modules take atom coordinates for each frame as input: 3N x t, where 3N are the atomic coordinates in the three Cartesian coordinates (x, y and z) and t denotes the different conformations.

SD2 module performs Principal Component Analysis (diagonalization of the covariance matrix) so as to eliminate dominant second order spatial correlations.[144] Apart from the atom coordinates for the different conformations, SD2 requires, the subspace dimensionality (m) as input. m can be selected by examining the cumulative variance plots that this module yields. Thus, SD2 carries out PCA, returning the eigenvalues (size m × 1), eigenvectors B (3N or D × m) and a projection matrix $Y = B^T X$ (m × t).[144,133]

SD4 module resolves the intrinsic non-orthogonal dependencies in positional fluctuations. The projection matrix, Y, from SD2 is used to build a fourth order spatially correlated cumulant tensor. Thus, SD4 diagonalizes this tensor and produces an anharmonic mode matrix W (3N or D × m). ANCA modes are thus ordered based on the kurtosis of the

projected coordinates; nevertheless, this ordering may not always correspond to a relevant reaction coordinate. To solve this, the user can define physical observables, more functionally and biologically relevant, e.g. the distant between two atoms or internal energy of the conformations.

TD2 module performs TICA (Time-lagged Independent Component Analysis) so as to remove dominant second order temporal correlations by computing a time-delayed covariance matrix. This module requires the atomic coordinates, and the subspace dimensionality of SD2 as input, but also another parameter the lag time (τ) over which the temporal correlations are to be resolved. This module returns Z, a matrix obtained by projecting the simulation data on the dominant eigenvectors and the eigenvalues.[151,152]

TD4 module is the temporal analogue of the spatial SD4 module, it builds a time-delayed fourth-order kurtosis tensor, which is then diagonalized to obtain anharmonic modes of fluctuations once the second order spatial and temporal correlations are resolved.[153] This module takes the matrix Z (from the TD2) as input, a user specified subspace value m denoting the number of desired anharmonic modes of motion, the lag time τ and the matrix V. The module returns the separating matrix W.

The outputs from the module TD4 can be used to build a MSM that summarise the transition pathway between substates.

4.1.2. Markov State Models (MSMs)

The essence of Markov state models (MSMs) is to construct a model with a series of discrete states and to parameterize the model with the inter-conversion rates between those states.[140] It is difficult to extract statistically relevant kinetic information by sheer visualization of MD trajectories. MSMs contribute to analyse a data set in a kinetically relevant manner.[140,147] By partitioning the conformational space into discrete states, MSMs create coarse models constituted of many states that results in building of a high-resolution model of intrinsic kinetics. In order to create a MSM one needs to gather the kinetically relevant structure states and transition rates between those structures. [145]

The first step in the creation of a MSM is a dimension reduction of the multidimensional conformational space. The dimension reduction is carried out through linear transformation methods such as Principal Component Analysis (PCA) or Time-lagged independent component analysis (TICA).[144,151] Trajectories are filtered through the independent and/or principal components.[151] Then, for the information to be gathered in a kinetic relevant manner, geometrical clustering methods such as k-means or k-centers are used.[154] This clustering would result in generation of many microstates structurally similar, which suggests a high level of kinetic similarity. To identify the kinetic relation between the microstates, it is necessary to construct a transition matrix, detailing the transition rates between microstates at a fixed time (lag time).[140,145]

To generate a transition matrix, the conformation of each frame of the trajectory is assigned to a microstate. Each structure in the trajectory is compared to the microstates. The closest microstate is identified and that structure is assigned to a relevant microstate. This means that the trajectories, which are frames over time, are converted to microstates.[145] The next step is to identify the transition rates between each pair of microstates i and j at a specific lag time. For instance, if a trajectory is at microstate i at time t then we want to know how many times the simulation entered the state j at time t+x. This is called the count matrix $C_{ij}()$. From that point, the probability of moving from i to j can be calculated in time which is known as $P_{ij}()$.[155,156] The probability of inter-conversion between two states is calculated within the mathematical framework of the Transition Path Theory (TPT).[157] The committor probability is essential to compute the transition pathway. The committor gives the probability to go from one intermediate state to the next one and not to the former.[140] Thus, TPT allows the conversion flux to be computed and provides the probability of any conversion way. On the other hand, sampling is an important element to construct a converged MSM.[145,158] Each microstate is not connected to the other, so to consider the number of transitions and amount of simulation per transition it is necessary to perform an efficient sampling.[145,146]

After the sampling of the microstate transition matrix, the MSM can be constructed.[140] Defining states in a ‘‘kinetically relevant” way requires that structures within a state can

interconvert on timescales faster than the lag time. The number of microstates can be reduced by using longer lag times. So, increasing the lag time means that states can get larger and more coarse grained. A coarser model is more easily understandable if any relevant intermediate state is lumped. Typically, the coarse graining of the states is done via some sort of spectral clustering of the microstate transition matrix. This is done by checking the eigenvalues and eigenvectors of the transition matrix to identify kinetically similar states. These allow one to define a coarse grained model at arbitrary resolution (high or low) depending on the goals for the model by lumping together kinetically related microstates.[140,145,146]

In this chapter, we present a MSM in which the dimensional reduction has been carried out using ANCA, instead of PCA or TICA.[145] The rest of the MSM has been constructed as conventional.

4.2. Methodology

The four simulations of the wild type protein in complex with Sap-C (3a and 3b) and alone (2a and 2b) were used to construct a kinetic model. Only the Cα coordinates of the residues belonging to the active pocket, were used in this model. The residues selected were: 120-130, 176- 181, 225- 261, 275- 296, 305-321, 340-356, 379-383 and 390-406. The analysis started from 400 ns onwards, considering the part before the equilibration time. Each original simulation consisted of 25000 frames (25 frames per ns) and was decreased to 5000 (skipping 5 frames), in order to accelerate the calculations. Thus, 3000 frames of each trajectory were processed in this analysis.

We used mainly two python libraries to carry out the analysis, namely PyAnca[143] and PyEmma[145]. Both were run in a Jupyter notebook[159]. Some complementary python libraries were implemented, intrinsically or by us, specifically: numpy[112], scipy[112], os[160], matplotlib[113], mdtraj[111], mdanalysis[161] and scikit-learn[162].

Trajectory and coordinates object were created using mdanalysis and mdtraj. A rigid body algorithm implemented in PyAnca: IterativeMeansAlign was used to align each step with the former one. Instant kurtosis, overall kurtosis and cumulative variance were obtained using scipy.stats (statistic module of scipy). RMSF and percentage anharmonicity for each residue were obtained within PyAnca.

4.2.1. Spatio-temporal decorrelation

Firstly, the module SD2 of PyAnca was used for removing spatial correlations. SD2 computes the covariance matrix and performs PCA, which decorrelates the factors with lag spacing of zero.[143] The dimension of the subspace was picked to be 60, based on the cumulative variance.

The module TD2 of PyAnca computes time-delayed covariance matrix and performs PCA, thus removing second order temporal correlations.[143] The dimensionality of the subspace

was specified to be 60, a decision based on the cumulative variance. The lag time selected was 1500 frames (300 ns), a time at which the different components stabilised.

Finally, the module TD4 of PyAnca was used for removing fourth order temporal correlations. TD4 performs a joint diagonalization of time-delayed cumulant matrices. The dimensionality of the subspace was selected to be 25, in order to accelerate the calculations. The lag time specified was 1500 frames (300 ns), the time at which the different components stabilised.

4.2.2. Markov State Model

The results from TD2 and TD4 coordinates were used to construct a MSM. First, the new coordinates were clustered using K-means algorithm within Pyemma[145]. The number of clusters selected was 250, which resulted in the formation of 250 microstates. A microstate is a set of different conformation that are structurally and energetically related. Then, the implied timescales were calculated for different lag times: 1, 2, 5, 10, 20, 50, 100, 200, 500, 800, 1000, and 2000. This served us to decide the lag time to use to estimate our MSM. The relaxation time of the slowest process (800 steps) was selected to calculate the model.

The Robust Perron Cluster Analysis Algorithm (PCCA) was used to coarse grain the data into a number of macrostates.[153] PCCA assigns a probability of membership of each microstate (generated by clustering) to a macrostate. Thus, each macrostate is constructed by lumping together smaller sets of energetically and structurally similar conformations called microstates. In our model we select 5 macrostates, a decision based on the relative relaxation timescales.[153,155] We finally obtained an MSM consisting of 5 macrostates as flux from a state A to a state B, with three intermediate states 0, 1 and 2.

The 50 conformation samples from PCCA were analysed for each state. The majoritarian conformations within the sample have been shown for each state. The visualization tool we used for the trajectories was VMD. [114]

4.3. Results

4.3.1. Anharmonic Conformational Analysis

In order to find the intrinsic dimensionality of the system, PCA was performed. PCA provides insights into how many modes are essential to estimate the number of conformational substates within the simulations. In our system, we identified that the first 10 eigenvalues accumulated the 80 % of the covariance and that 60 eigenvalues accounted for the 95 % on the covariance (Fig. 4.1). This means that our system has 60 intrinsic dimensions, which the conformations will be projected on. The 60 dimensional space is also the space in which SD2 and TD2 calculation will be carried out.

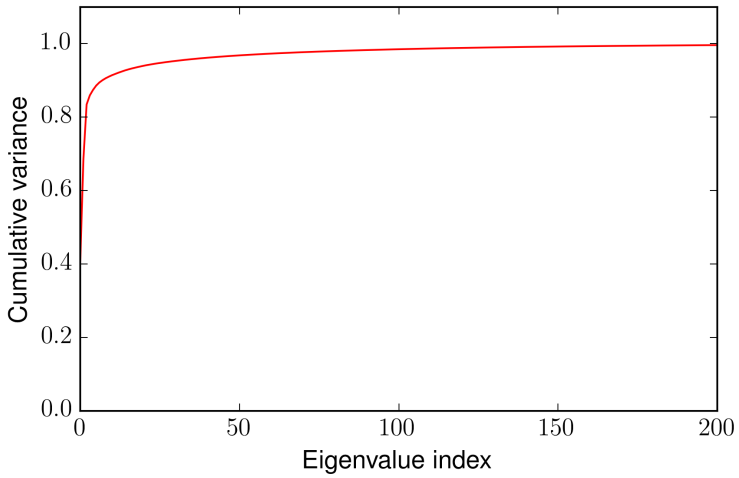

Figure 4.1: *Cumulative variance of the system. The first 60 eigenvalues account for 95 % of the variance. This means that 60 is the intrinsic dimensionality of our system - the space that will be used for future calculations.*

Kurtosis was used to quantify anharmonicity (non-Gaussianity) from positional deviations. Non-Gaussian atomic deviations are associated with functional regions of the protein.[134] The overall kurtosis for our system was 5.41, so that it presents a super-Gaussian ($\kappa > 3$) distribution, and has a median of 4.5. Super-Gaussian distributions (G^S) are more picked and heavier tailed than Gaussian distributions as can be seen in Figure 4.2 (b).

We have calculated what parts of the active site exhibit anharmonic motions and for how long. As we just selected the active site pocket for analysis. Most of the parts analysed exhibited high anharmonicity as illustrated in the figures 4.2(a) and 4.2(c). Loops-1 and -5 exhibit high percentage of anharmonicity, along with the regions surrounding them. Loops-2, -3 and -4 exhibit less percentage of anharmonicity. Regarding the kurtosis of individual residues, Loop-5 and surroundings exhibited high kurtosis, whereas the rest of the loops exhibited similar values. We have also included the values of RMSF per residue to compare with these results (Figure 4.2(d)). In the RMSF values fast motions have not been removed from the calculations, which is the reason why both measures RMSF and Kurtosis present different results.

Spatial decorrelation was carried out using SD2 and selecting a subspace of 60 dimensions. The coordinates from SD2 (matrix Y) was used as input for TD2, which performed second order temporal decorrelation. The coordinates from SD2 were then used as input for TD2. A lag time of 1500 frames was selected to run TD2. Finally, the output from TD2 (coordinates in matrix called Z) were used as input for TD4. TD4 carries out fourth order temporal decorrelation. TD4 was run using a lag time of 1500 and a space dimensionality of 25 dimensions.

Figure 4.3. shows the trajectory through the first four components of TD4. As it is a time-lag delayed calculation we called those components (TICA). It can be seen in the figure how TD4 has reduced the trajectories to a succession of discrete jumps as a result of dimensional reduction. The coordinates of TD4 were clustered using K-means algorithm. In order to improve the visualization, a histogram has been built of the first two TD4 dimensions and the free energy computed. Figure 4.4(a) shows this histogram and 4.4(b) shows the clusters over the histogram. Each one of the clusters represents a microstate. Later in the analysis, the microstates will become a part of a larger macrostate, and the system will be reduced to a few states.

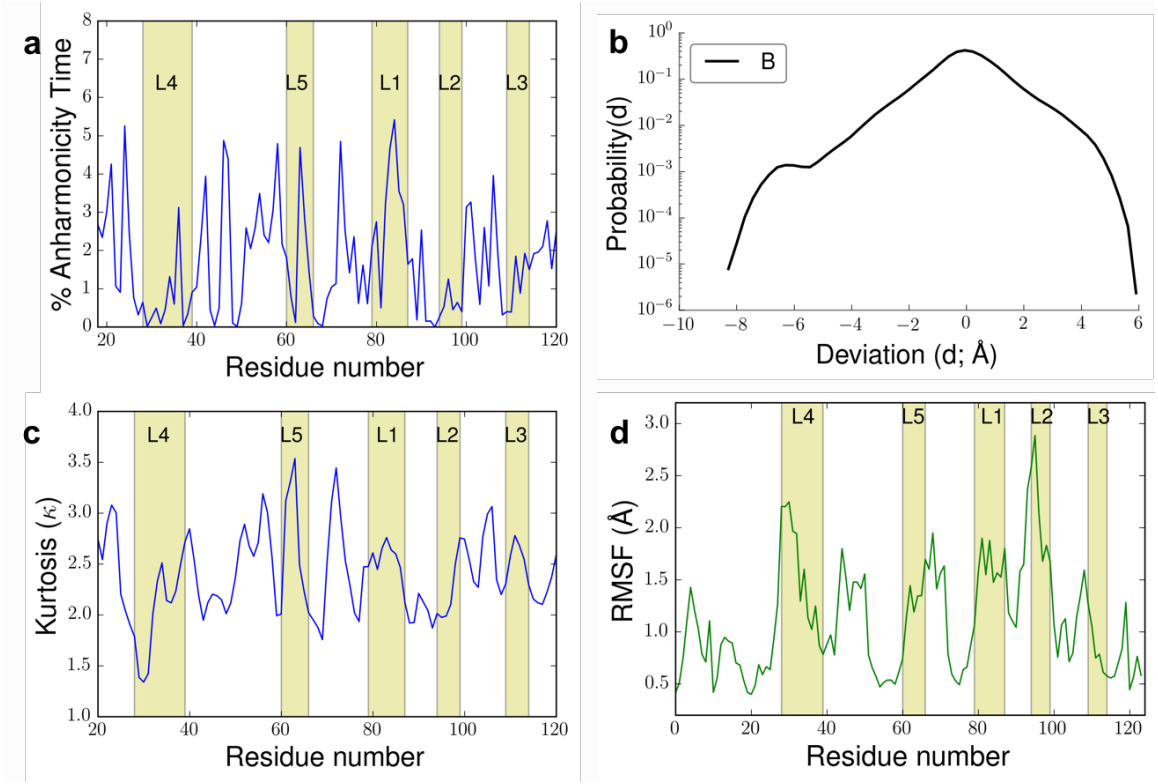

Figure 4.2. *(a) Percentage of anharmonic time in our system. Loop-1 and Loop-5 exhibit higher percentage of anharmonicity. (b) Overall kurtosis of the system, illustrated as a super Gaussian distribution with a median value of 5.41. Positional desviations of the Cα are anharmonic, non-Gaussian. (c) Kurtosis values per residue: Loop-5 exhibits the greatest value of kurtosis, while the values of Loop-1 and Loop-3 were higher than the values of Loop-4 and Loop-5. (d) RMSF values per residue. Please note that the higher RMSD fluctuations do not involve high anharmonic percentage time or value of kurtosis per residues (fast motions have not been removed from the calculations).*

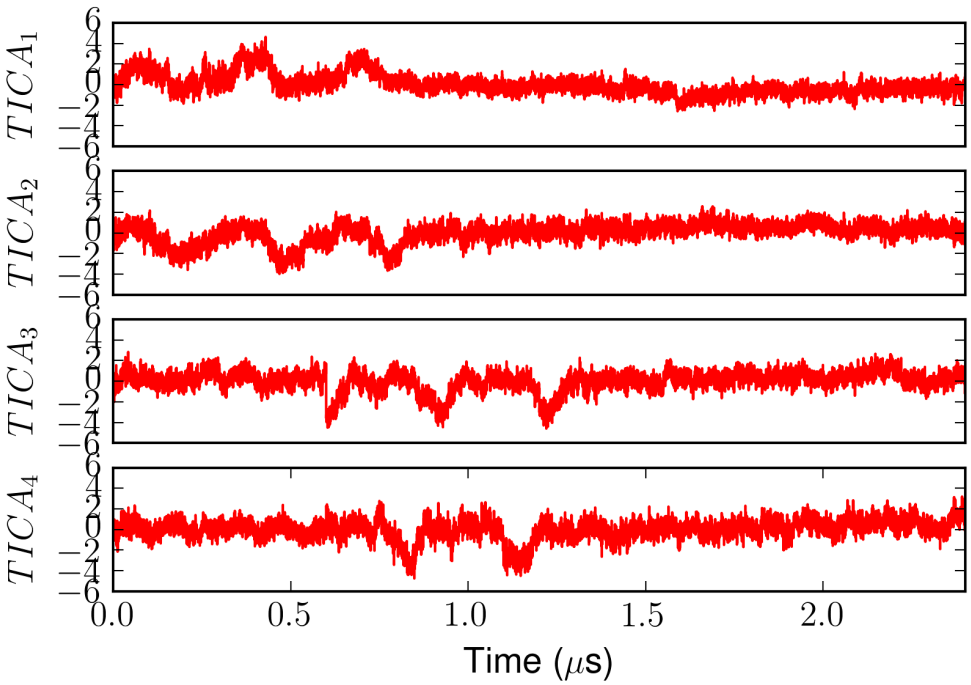

Figure 4.3: *Trajectory filtered through the first four TD4 components. Temporal and spatial decorrelation by modules SD2 and TD2 and TD4 facilitates the extraction of anharmonic modes of motion depriving the trajectory of Gaussian noise. The trajectory is thus reduced to a small set of discrete jumps, projected onto TD4 component values, that represent the independent, anharmonically fluctuating, protein motions.*

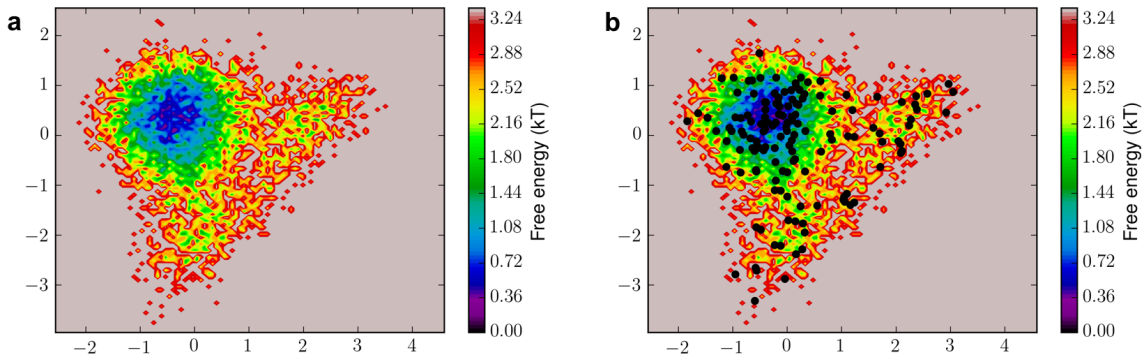

Figure 4.4: *(a and b) Histogram of the first two TD4 components TICA 1 (x-axis) and TICA2 (y-axis) and the computed free energy (b) Clusters as obtained from k-means are showed as black dots.*

4.3.2. Markov State Model

In this step, appropriate lag time selection as well as space discretization is performed. To identify the lag time, different lag times were scanned and relaxation timescale computed for each MSM. Initially the timescales are dependent of the lag times but after 800 steps (160 ns), it seems to stabilize (Figure 4.5). Thus a lag time of 800 steps was selected for the generation of a MSM using the coordinates from TD4.

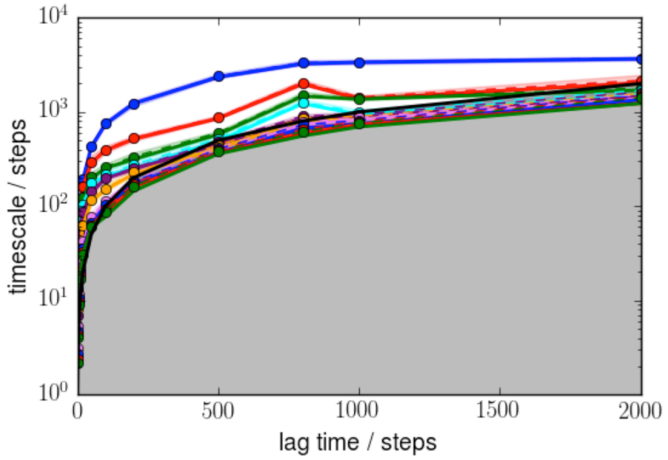

Figure 4.5*: Relaxation timescales of different MSMs at different lag times. At small lag times the relaxation time is dependent on the lag time. From 800 steps onwards, the relaxation time of the slowest process is constant. The grey area indicates that the relaxation time is smaller than the lag time so the model cannot be predicted.*

Furthermore, spectral analysis was performed to obtain the calculated eigenvectors and identify the dominant motions in the TD4 analysis. The eigenvalues were represented by dots and the plots represent the overall impact of the 10 dominant eigenvectors in the overall motion (Figure 4.6(a)). This plot allows us to see the relaxation timescale of the dominant motions. Taking the ratios of the different relaxation timescales permits us to see the separation times between the different processes (Figure 4.6(b)). Relaxation timescales equates in faster motions, thus we can represent a model just by retaining those relaxation timescales that present more differences. In our model, we see how there is a gap between timescale separation of eigenvectors 0 and 1 and between 2 and 3 and 4, so that in our case it would be a good model if we choose 5 macrostates.

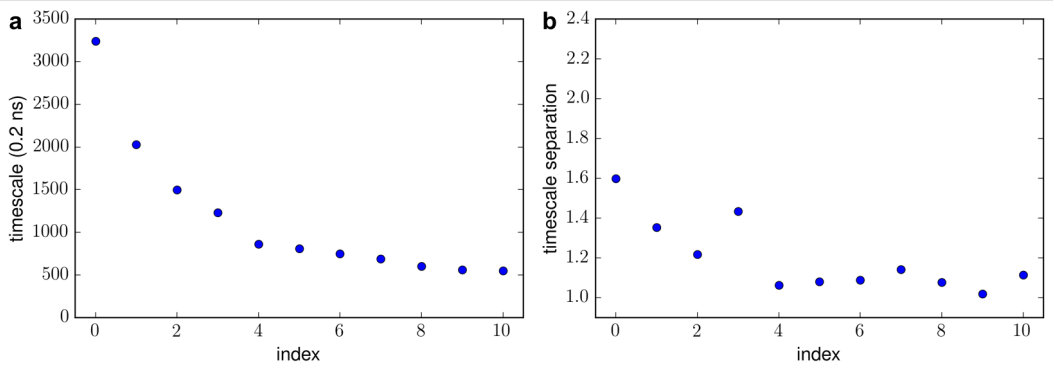

Figure 4.6: *(a) Relaxation timescales of the dominant motions. (b) Relaxation timescale separations for the different process. Those with higher separation will be taken to represent the model.*

MSM's as well as other kinetic models are approximations and therefore have non-zero systematic error. This error will depend on the type of model used, the lag time selected and the state space discretization. It is, therefore, mandatory to validate the kinetic model before using it for analysis.[145] When the model, at a lag time τ , is capable of predicting estimates performed at a longer time scales $\kappa\tau$ within statistical error, the validation is considered successful. Chapman-Kolmogorov test compares the prediction and the estimation of probability of being in a set J at time $\kappa\tau$. The quality of our model was assessed using Chapman-Kolmogorov test. Our model seemed to adjust to the estimation, Figure 4.7.

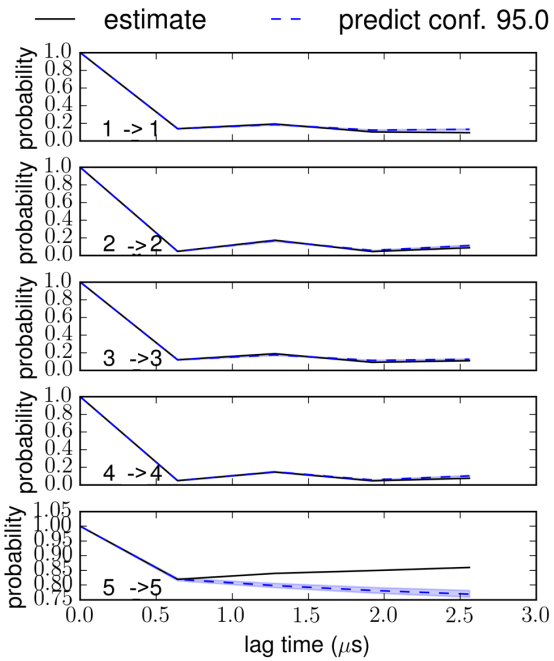

Figure 4.7. *Quality of the model as assessed by Chapman-Kolmogorov test. The model is depicted as a black continuous line whereas the estimation is depicted as blue hashed line. It can be appreciated that our model correlated well to the estimation.*

The MSM microstates extracted by clustering ranges to hundreds and in some cases to thousands, do not provide a human understandable system. A coarse-grained model containing macrostates structures is generated by grouping the structures and adding up their equilibrium probabilities.[156,163]

Continuous molecular processes contain transition states that are not easily assignable to a cluster. To overcome the problem of transition states the concept of 'fuzzy clustering' was introduced. 'Fuzzy clustering' assigns every object to all the clusters with certain probability. The coarse grain process is done through PCCA analysis using fuzzy clustering. PCCA determines the probability for each microstate to belong to a given macrostate. The probabilities are called memberships to a given macrostate. In the Figure 4.8. we have plotted the Bayesian inverse. This is the probability of the structures that remain as a microstate when it is in a metastable state.

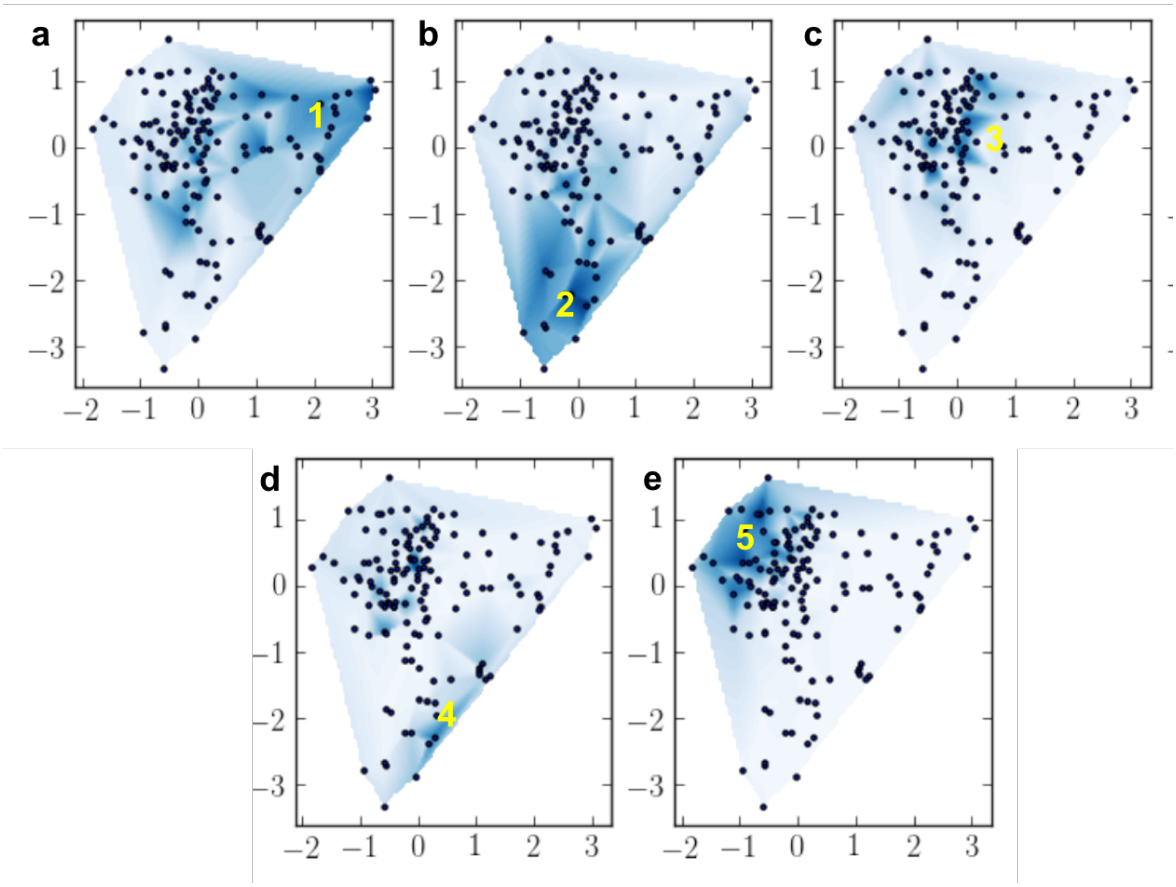

Figure 4.8: (a-e) *Bayesian inverse plots for the distribution of the five longest living metastable states. The blue shadow represents the probability of membership to a metastable state, the darker the blue is the higher is the probability of belonging to a determinate metastable state. The probality is calculated for the first five components and projected in five plots in which TICA 1 (x-axis) and TICA2 (y-axis) are the axis. The predicted metastable states have been numbered in yellow.*

Coarse-graining further filtered irrelevant information. However, to construct a detailed markov state model, a transition pathway needs to be generated. In order to create a comprehensive transition pathway system, another PCCA clustering is employed targeting 5 metastable states.

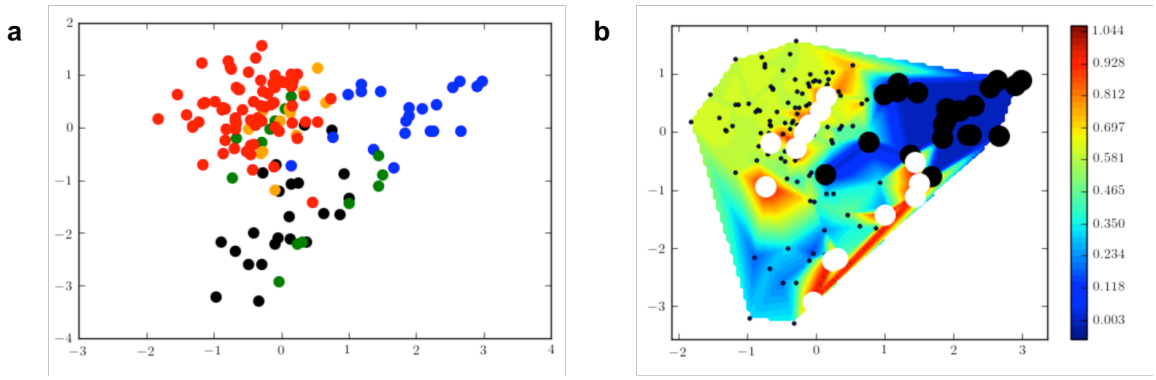

Figure 4.9: *(a) Predicted metastable states projected on the first two eigenvectors. Each microstate has been coloured differently depending of the membership to a metastable state. (b) Demonstrates the two end states of the model within the density plot where the blue represents high density regions and the red shows the low density. Black dots are microstates belonging to the macrostate A whereas white dots are those belonging to the macrostate B.*

Moreover, the estimated time to go from A macrostate to B is 4.3 microseconds and 6.1 to come back.

We finally plot our model with the transition probabilities (Figure 4.10). The probability fluxes of the pathway in the model are from the rightmost state to the leftmost state. In the model; the pathways moves from one stable state and then splits into two basic intermediary states, 0 and 2. From 0 it goes either to 2 or straight to B. From 2 it goes to an intermediate state 1 and from there to B or straight to B. 20 samples of each macrostate were taken and the 10 more similar were chosen to represent the macrostate.

Figure 4.10 shows our final MSM. The model is composed of five macrostates as explained above. We have taken 50 samples from the PCCA for sampling each state. The majoritarian conformation among those 50 sample has been chosen to represent the state.

The state A is represented by a structure that has Loop-1 in extended conformation while Loop-2 and Loop-3 have a closed conformation. From state A the fluxes split in four to state 0 or state 2 with high probability or to State 1 and B with little probability.

State 0 is represented by a partially active conformation of the enzyme. Loop-1 is helical although partially extended, Loop-2 is in closed conformation and Loop-3 is in open conformation. From this state the flux goes to state 2.

The set of representative structures of state 1 exhibit an extended conformation of Loop-1, extended towards the helix 7, and Loop-2 and -3 in closed conformation. From State 1 the flux goes to state B.

The set of representative structures in state 2 are characterised by an extended conformation of Loop-1, and open conformation of Loops-2 and -3. From State 2 it goes either to State 1 or B.

Finally, State B is represented by an active conformation of GCase that has a helical conformation of Loop-1, and open conformation of Loops 2 and 3.

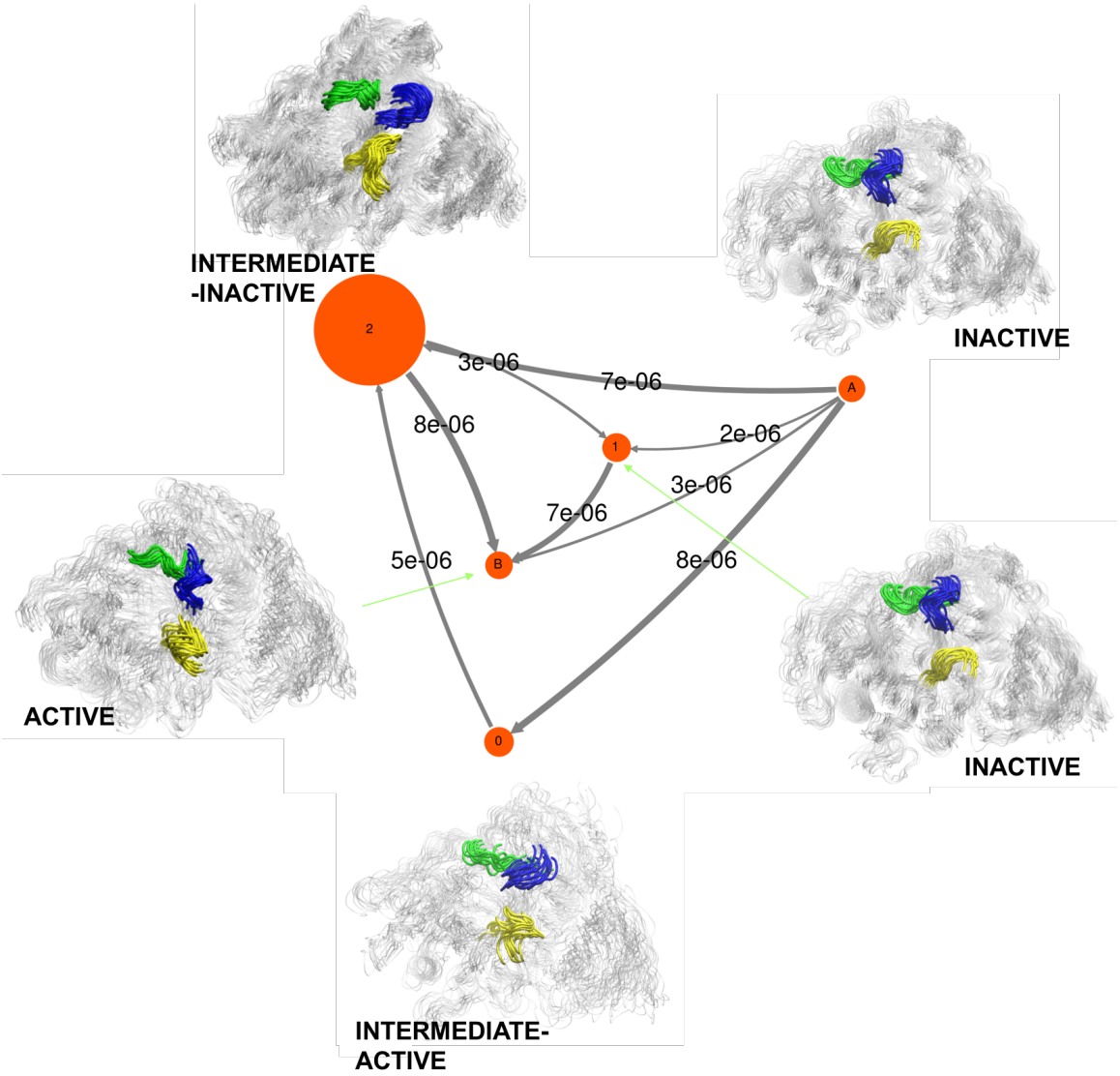

Figure 4.10: *The Markov State Model is composed of 5 macrostates. The set of structures, representative of each state has been illustrated in white with loops-1, -2 and -3 coloured in yellow, blue and green respectively. Inactive, intermediates (active and inactive) and active conformations of the protein has been labelled. The numbers are the transition times expressed in seconds.*

4.4. Discussion

Atomistic MD is still not sufficient to draw a complete picture of the FES of a protein.[141] The traditional MD analysis that measure conformational drifts, such as RMSD or radius of gyration, cannot be used to infer dominant motions accountable for protein function.[140] In this Chapter, higher order statistics and Markov state models have been combined to construct a kinetic model that allows us to dissect the conformational FES of the GCase, providing insights into the activation process.

The analysis has been carried out exclusively on the binding pocket of GCase to build this kinetic model. A total of 12000 conformations from 4 different simulations of the wild type of the enzyme GCase have been analysed. The quantification of anharmonicity using kurtosis as a measure, is able to recognise the functional parts of the proteins. Kurtosis and percentage of anharmonicity indicated Loop-1 as functionally relevant, whereas secondary role of the other Loops was also detected. However, it must be stressed here that only Cα atoms have been used in the analysis. This is in contrast to classical simulations where the open and closed state of Loop-2 and Loop-3 were influenced by sidechains positioning of residues W348 in Loop-2 and R395 in Loop-3, as explained previously.

As the meaningful motions of the protein are concentrated in just a few dimensions it is unnecessary to cope with a multidimensional complexity.[144,145,158] Fourth order statistics has been used to reduce the dimension of the data. TD4 was the final step in a series of transformations of our data. TD4 has been able to reduce our trajectory to a collection of discrete jumps. The components of TD4 can be further processed to isolate relevant discrete states liable for the functionality of the protein.

The construction of a MSM is an efficient way to extract information from the eigenvectors and dispose it in a human readable format.[140] The relaxation time of the slowest motion was around 800 ns. This suggests that transitions faster than 800 ns will be ignored in the results. We decided to represent the model in 5 substates after the relaxation timescale separation.

The MSM shows the transition among states. Our model describes a transition path in which the enzyme goes form an initial state A to a final state B via 3 different substates. During this transition we identified inactive, intermediate and active conformations of the protein. State A is represented by an inactive conformation of the protein, characterised by an extended conformation of Loop-1 and Loop-2 and -3 oriented towards the binding site. State B is represented by an active conformation of the enzyme characterised by a helical conformation of Loop-1, and open conformations of Loop 2 and 3. From State A to B the enzyme goes through a series of intermediate states. State 0 is represented by a partial active form (Loop-1 helical), where some features of the inactive conformation are present. Figure 4.10 illustrates how Loop-2 and -3 are oriented towards the binding site. From 0 the flux goes to State 2. State 2 is the lowest energy state, and hence extensively populated. The conformation representative of this state is again an intermediate state. Loop-1 is in extended conformation, although Loop-2 is dramatically open towards helix 7 and Loop-3 is oriented toward the outside of the binding site. Finally, State 1 is a typical inactive conformation with an extended conformation of Loop-1. In this model the inactive enzyme passes through three intermediates states, to end up in an active conformation. It is important to note that the marked low energy state (State 2) does not correspond to any of the crystal structures. However, low energetic conformations are usually highly populated and can be easily isolated by physical techniques.[135] It is important to stress here that the conformation in State 2 is similar to that observed towards the end of simulation 4b (inactive GCase and Sap-C), where Loop-2 was inserted in a hydrophobic pocket under Sap-C. The same conformation was proposed to influence Loop-3 to open towards the outside of the binding site and allow stable binding for the residue K33 of Sap-C to activate the Loop-1. It is possible that this very stable conformation has never been isolated because the structure of GCase along with Sap-C has never been studied physically.

CHAPTER 5:

CONCLUSIONS

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

Understanding functional mechanism of the enzyme GCase would shed light on new ways to address the treatment for Gaucher Disease. In this thesis, I have presented a protein-protein interaction model of the GCase with its facilitator protein, Sap-C, anchored into a lipid bilayer using Coarse- Grained molecular dynamics. The details of the interactions between different components and their evolution over time were observed by carrying out Atomistic Molecular Dynamics simulations. The difference in dynamics between wild type proteins and the clinically most important mutants (N370S and L444P) have also been addressed. A kinetic model accounting for the activation mechanisim of the enzyme has been presented.

Protein- protein model

The protein-protein interaction model proposed here is in agreement with experimental studies GCase-Sap-C binding. This model is also coherent with membrane anchoring experiments of both proteins separately and correlates well with the recent crystal structure of related GalC and Sap-A co-complex. The model was obtained using two different docking techniques. The model passed all the filters that we constructed to find our favourable conformation.

Molecular Dynamics

The anchoring of the GCase models in a DPPC membrane was performed in 5 different simulation environments, including, with and without facilitator protein Sap-C, and with and without substrate GluCer. In all cases GCase, anchored to the membrane and we were able to analyse the differences. The mode of binding obtained is coherent with experimental studies and theoretical predictions of interation of the proteins with the membrane.

The analysis of Atomistic MD simulations provided the details of membrane anchoring, protein-protein and ligand binding. All systems showed overall structure stability. The main fluctuations were identified in external surface loops. The simulations including Sap-

C helped us to locate important hydrogen bond networks at the protein-protein interface.

In this thesis, I propose a mechanism of helication of Loop-1 through the interaction of residue D315 with residue K33 of Sap-C. This interaction was stable through the entire simulation in the active wild type simulation with the facilitator protein. The interaction was also observed in simulations of mutants containing Sap-C. The structural differences between these systems allowed us to study the details of this interaction. In simulation of the inactive wild type, residue K33 from Sap-C interacted with residues in the same Loop as D315, however this interaction was not observed in the mutants.

The interaction between K33 (Sap-C) and Loop-1 are observed in the wild type (active and inactive) but not in the mutant. This is primarily because of a large number of hydrophobic interactions that stabilize the protein-protein interface. The stabilization of residue W348 (Loop-2) in a hydrophobic pocket under Sap-C coincides with the disruption of interactions occurring between Loop-2 and Loop-3. The loss of these interactions allow Loop-3 to adopt an open conformation. Residue R395 of Loop-3 was found to make an ion-pair interactions with E340 in mutants and in GCase simulated without Sap-C. We observed that in mutants, W348 (Loop-2) is not stabilized in the hydrophobic pocket formed at the GCase-Sap-C interface.

Mutants that were constructed from the active conformation, begin to show some features of inactivation, as the simulation progresses such as desestabilization of the residue W348 in the hydrophobic pocket at the GCase-Sap-C interface and closure of Loop-3 in the case of L444P. Some of those inactive conformations features were also observed in simulation 2a where the protein is simulated without Sap-C.

Kinetic studies

In this thesis, I present a kinetic model of activation of the enzyme GCase. The model was built using anharmonic conformational analysis (ANCA) and markov state models (MSMs). ANCA was able to detect the relevant role of Loop-1 in the activation process that traditional measures such as RMSF were not able to define. ANCA was used to

perform Spatio-Temporal decorrelation, and thus reduce the dimension of space and its discretization. ANCA reduced our trajectories to a set of discrete jumps, which can be easily analysed.

An MSM model was built using the data reduced from ANCA. This model highlights an activation path of the enzyme GCase. The protein goes from an inactive state to an active state passing through three intermediate states. One of these intermediate states (2) was highly stable and extensively populated. The conformations that represent this State 2 were observed towards the end of simulation 4b. The intermediate state 2 can be correlated with the conformation that we proposed for the transition state structure, from active to inactive, observed in classical molecular dynamics.

FUTURE WORK

First, towards the end of the work that is presented here, the crystal structure of the enzyme GalC-Sap-A was published. Though belonging to the same family, GalC and GCase perform very different functions. The crystal structure shows Sap-A in an open conformation. Sap-C is only present in open conformation under detergent conditions. A dynamic model of GCase- open Sap-C has not been presented here. It should be simulated and included in future studies.

Second, the atomistic simulations should be extended in order to improve the sampling of the FES of our system. While our simulations have equilibrated, they have not yet converged at the current timescale. Convergence will help us observe the end of the activation process proposed in this thesis.

Third, a kinetic model with extended simulations could give us more information about the activation process including intermediates. Furthermore, a model that includes mutants can provide a more complete view of the FES of GCase.

Finally, and although ambitious, structural implications of every mutation identified in Gauchers diseases should be looked into. A datebase of mutations and their structural implications will permit us to create a model able to link structural information with

phenotypic consequences.

REFERENCES

1. Grabowski, G. A. Phenotype, diagnosis, and treatment of Gaucher's disease. *The Lancet* **372,** 1263–1271 (2008).

2. Sidransky, E. Gaucher Disease: Insights from a Rare Mendelian Disorder.

3. Hruska, K. S., LaMarca, M. E., Scott, C. R. & Sidransky, E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum. Mutat.* **29,** 567–83 (2008).

4. Lieberman, R. L. A Guided Tour of the Structural Biology of Gaucher Disease: Acid-β-Glucosidase and Saposin C. *Enzyme Research* (2011). doi:10.4061/2011/973231

5. Karplus, M. & Petsko, G. a. Molecular dynamics simulations in biology. *Nature* **347,** 631–639 (1990).

6. Kolter, T. & Sandhoff, K. Lysosomal degradation of membrane lipids. *FEBS Letters* **584,** 1700–1712 (2010).

7. Schulze, H., Kolter, T. & Sandhoff, K. Principles of lysosomal membrane degradation. Cellular topology and biochemistry of lysosomal lipid degradation. *Biochimica et Biophysica Acta - Molecular Cell Research* (2009). doi:10.1016/j.bbamcr.2008.09.020

8. Kolter, T. & Sandhoff, K. Lysosomal degradation of membrane lipids. *FEBS Letters* (2010). doi:10.1016/j.febslet.2009.10.021

9. Hannun, Y. A. & Obeid, L. M. Principles of bioactive lipid signalling: lessons from sphingolipids. *Nat. Rev. Mol. Cell Biol.* **9,** 139–150 (2008).

10. Schultz, M. L., Tecedor, L., Chang, M. & Davidson, B. L. Clarifying lysosomal storage diseases. *Trends in Neurosciences* **34,** 401–410 (2011).

11. Nagral, A. Gaucher disease. *Journal of Clinical and Experimental Hepatology* **4,** 37–50 (2014).

12. Mistry, P. K. *et al.* Glucocerebrosidase gene-deficient mouse recapitulates Gaucher disease displaying cellular and molecular dysregulation beyond the macrophage. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.1003308107

13. Liu, J. *et al.* Gaucher disease gene GBA functions in immune regulation. *Proceedings of the National Academy of Sciences* (2012). doi:10.1073/pnas.1200941109

14. Lwin, A., Orvisky, E., Goker-Alpan, O., LaMarca, M. E. & Sidransky, E. Glucocerebrosidase mutations in subjects with parkinsonism. *Mol. Genet. Metab.* **81,** 70–73 (2004).

15. Tayebi, N. *et al. Gaucher disease and parkinsonism: a phenotypic and genotypic characterization. Molecular genetics and metabolism* **73,** 313–321 (2001).

16. Clark, L. N. *et al.* Mutations in the glucocerebrosidase gene are associated with early-onset Parkinson disease. *Neurology* **69,** 1270–1277 (2007).

17. Shayman, J. A., Abe, A. & Hiraoka, M. A turn in the road: How studies on the pharmacology of glucosylceramide synthase inhibitors led to the identification of a lysosomal phospholipase A2 with ceramide transacylase activity. in *Glycoconjugate Journal* **20,** 25–32 (2004).

18. Lee, L., Abe, A. & Shayman, J. A. Improved inhibitors of glucosylceramide synthase. *J. Biol. Chem.* **274,** 14662–14669 (1999).

19. Pastores, G. M., Giraldo, P., Chérin, P. & Mehta, A. Goal-oriented therapy with miglustat in Gaucher disease. *Curr. Med. Res. Opin.* **25,** 23–37 (2009).

20. Reczek, D. *et al.* LIMP-2 Is a Receptor for Lysosomal Mannose-6-Phosphate-Independent Targeting of β-Glucocerebrosidase. *Cell* **131,** 770–783 (2007).

21. Zachos, C., Blanz, J., Saftig, P. & Schwake, M. A Critical Histidine Residue Within LIMP-2 Mediates pH Sensitive Binding to Its Ligand ??-Glucocerebrosidase. *Traffic* **13,** 1113–1123 (2012).

22. Berent, S. L. & Radin, N. S. Mechanism of activation of glucocerebrosidase by co-beta-glucosidase (glucosidase activator protein). *Biochim. Biophys. Acta* **664,** 572–582 (1981).

23. Sun, Y., Qi, X. & Grabowski, G. A. Saposin C is required for normal resistance of acid β-glucosidase to proteolytic degradation. *J. Biol. Chem.* (2003). doi:10.1074/jbc.M302752200

24. Vasella, A., Davies, G. J. & Böhm, M. Glycosidase mechanisms. *Current Opinion in Chemical Biology* **6,** 619–629 (2002).

25. Dvir, H. *et al.* X-ray structure of human acid-β-glucosidase, the defective enzyme in Gaucher disease. *EMBO Rep.* **4,** 704–709 (2003).

26. Henrissat, B. & Bairoch, A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **293 (Pt 3,** 781–788 (1993).

27. Lieberman, R. L. *et al.* Structure of acid beta-glucosidase with pharmacological chaperone provides insight into Gaucher disease. *Nat. Chem. Biol.* **3,** 101–107 (2007).

28. Wei, R. R. *et al.* X-ray and biochemical analysis of N370S mutant human acid ??-glucosidase. *J. Biol. Chem.* **286,** 299–308 (2011).

29. Kacher, Y. *et al.* Acid beta-glucosidase: insights from structural analysis and relevance to Gaucher disease therapy. *Biol. Chem.* **389,** 1361–1369 (2008).

30. Berg-Fussman, A., Grace, M. E., Ioannou, Y. & Grabowski, G. A. Human acid β-glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. *J. Biol. Chem.* **268,** 14861–14866 (1993).

31. Salvioli, R. *et al.* The N370S (Asn370-->Ser) mutation affects the capacity of glucosylceramidase to interact with anionic phospholipid-containing membranes and saposin C. *Biochem. J.* **390,** 95–103 (2005).

32. Colombo, R. Age estimate of the N370S mutation causing Gaucher disease in Ashkenazi Jews and European populations: A reappraisal of haplotype data. *Am. J. Hum. Genet.* **66,** 692–697 (2000).

33. Pasmanik-Chor, M. *et al.* The glucocerebrosidase D409H mutation in Gaucher disease. *Biochem. Mol. Med.* **59,** 125–133 (1996).

34. Bendikov-Bar, I., Ron, I., Filocamo, M. & Horowitz, M. Characterization of the ERAD process of the L444P mutant glucocerebrosidase variant. *Blood Cells, Mol. Dis.* **46,** 4–10 (2011).

35. Sun, Q. Y. *et al.* Glucocerebrosidase gene L444P mutation is a risk factor for Parkinson's disease in chinese population. *Mov. Disord.* **25,** 1005–1011 (2010).

36. Horowitz, M., Pasmanik-Chor, M., Ron, I. & Kolodny, E. H. The enigma of the E326K mutation in acid ??-glucocerebrosidase. *Molecular Genetics and Metabolism* **104,** 35–38 (2011).

37. Duran, R. *et al.* The glucocerobrosidase E326K variant predisposes to Parkinson's disease, but does not cause Gaucher's disease. *Mov. Disord.* **28,** 232–236 (2013).

38. Ahn, V. E., Leyko, P., Alattia, J.-R., Chen, L. & Privé, G. G. Crystal structures of saposins A and C. *Protein Sci.* **15,** 1849–57 (2006).

39. Fabbro, D. & Grabowski, G. A. Human acid β -glucosidase. Use of inhibitory and activating monoclonal antibodies to investigate the enzyme's catalytic mechanism and saposin A and C binding sites. *J. Biol. Chem*. **266,** 15021-27 (1991).

40. Tamargo, R. J., Velayati, A., Goldin, E. & Sidransky, E. The role of saposin C in Gaucher disease. *Molecular Genetics and Metabolism* **106,** 257–263 (2012).

41. Tylki-Szymańska, A. *et al.* Gaucher disease due to saposin C deficiency, previously described as non-neuronopathic form - No positive effects after 2-years of miglustat therapy. *Mol. Genet. Metab.* **104,** 627–630 (2011).

42. Salvioli, R. *et al.* The N370S (Asn370-->Ser) mutation affects the capacity of glucosylceramidase to interact with anionic phospholipid-containing membranes and saposin C. *Biochem. J.* (2005). doi:10.1042/BJ20050325

43. Qi, X., Qin, W., Sun, Y., Kondoh, K. & Grabowski, G. A. Functional organization of saposin C: Definition of the neurotrophic and acid ??-glucosidase activation regions. *J. Biol. Chem.* **271,** 6874–6880 (1996).

44. Weiler, S., Kishimoto, Y., O'Brien, J. S., Barranger, J. A. & Tomich, J. M. Identification of the binding and activating sites of the sphingolipid activator protein, saposin C, with glucocerebrosidase. *Protein Sci.* **4,** 756–64 (1995).

45. Weiler, S., Kishimoto, Y., O'Brien, J. S., Barranger, J. A. & Tomich, J. M. Identification of the binding and activating sites of the sphingolipid activator protein, saposin C, with glucocerebrosidase. *Protein Sci.* **4,** 756–764 (1995).

46. Atrian, S. *et al.* An evolutionary and structure-based docking model for glucocerebrosidase- saposin C and glucocerebrosidase-substrate interactions - Relevance for Gaucher disease. *Proteins Struct. Funct. Genet.* (2008). doi:10.1002/prot.21554

47. Zubrzycki, I. Z., Borcz, A., Wiacek, M. & Hagner, W. The studies on substrate, product and inhibitor binding to a wild-type and neuronopathic form of human acid- β -glucosidase. *J. Mol. Model.* (2007). doi:10.1007/s00894-007-0232-5

48. Offman, M. N., Krol, M., Silman, I., Sussman, J. L. & Futerman, A. H. Molecular basis of reduced glucosylceramidase activity in the most common Gaucher disease mutant, N370S. *J. Biol. Chem.* **285,** 42105–42114 (2010).

49. Jones, G., Willett, P. & Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245,** 43–53 (1995).

50. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided. Drug Des.* **7,** 146–157 (2011).

51. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins Struct. Funct. Genet.* **47,** 409–443 (2002).

52. Huang, S.-Y. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov. Today* **19,** 1081–1096 (2014).

53. Ritchie, D. W. Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.* **9,** 1–15 (2008).

54. Gabb, H. a, Jackson, R. M. & Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272,** 106–120 (1997).

55. Ritchie, D. W. Evaluation of Protein Docking Predictions Using Hex 3 . 1 in CAPRI Rounds 1 – 2. *Evaluation* (2003).

56. Strynadka, N. C. *et al.* Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat. Struct. Biol.* **3,** 233–239 (1996).

57. Ritchie, D. W. & Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Genet.* (2000). doi:10.1002/(SICI)1097-0134(20000501)39:2<178::AID-PROT8>3.0.CO;2-6

58. Damm, W., Frontera, A., Tirado–Rives, J. & Jorgensen, W. L. OPLS all-atom force field for carbohydrates. *J. Comput. Chem.* **18,** 1955–1970 (1997).

59. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125,** 1731–1737 (2003).

60. de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5,** 883–897 (2010).

61. Huang, B. & Schroeder, M. Using protein binding site prediction to improve protein docking. *Gene* **422,** 14–21 (2008).

62. de Vries, S. J. & Bonvin, A. M. J. J. Cport: A consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* **6** (3)**:** e17695 (2011).

63. Neuvirth, H., Raz, R. & Schreiber, G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338,** 181–199 (2004).

64. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. in *Proteins: Structure, Function and Genetics* **66,** 630–645 (2007).

65. Qin, S. & Zhou, H.-X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* **23,** 3386–3387 (2007).

66. Liang, S., Zhang, C., Liu, S. & Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **34,** 3698–3707 (2006).

67. Ritchie, D. Hex 6.1 User Manual. *Molecules* (2010).

68. Chuang, G.-Y., Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys. J.* **95,** 4217–4227 (2008).

69. Rossmann, M. *et al.* Crystal Structures of Human Saposins C and D: Implications for Lipid Recognition and Membrane Interactions. *Structure* **16,** 809–817 (2008).

70. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3,** 198–210 (2013).

71. Hill, C. H. *et al.* The mechanism of glycosphingolipid degradation revealed by a GALC-SapA complex structure. *Nat. Commun.* **9,** articule number: 151 (2018).

72. Deane, J. E. *et al.* Insights into Krabbe disease from structures of galactocerebrosidase. *Proc. Natl. Acad. Sci.* **108,** 15169–15173 (2011).

73. Qi, X. & Grabowski, G. A. Differential membrane interactions of saposins A and C: Implications for the functional specificity. *J. Biol. Chem.* **276,** 27010–27017 (2001).

74. Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6679–6685 (2005).

75. Hospital, A., Goñi, J. R., Orozco, M. & Gelpi, J. Molecular dynamics simulations: Advances and applications. *Adv. Appl. Bioinforma. Chem.* **8,** 37–47 (2015).

76. Adcock, S. A. & McCammon, J. A. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews* **106,** 1589–1615 (2006).

77. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **41,** 429–452 (2012).

78. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of

biomolecules. *Nat. Struct. Biol.* **9,** 646–652 (2002).

79. Hehre, W. J. *A Guide to Molecular Mechanics and Quantum Chemical Calculations*. *Interpreting* (2003).

80. Born, M. & Oppenheimer, J. R. Born-Oppenheimer approximation. *Ann. Phys.* **84,** 457 (1927).

81. Doltsinis, N. L. Molecular Dynamics Beyond the Born-Oppenheimer Approximation : Mixed Quantum – Classical Approaches. *Comput. Nanosci. Do It Yourself!* **31,** 389–409 (2006).

82. Woolley, R. G. & Sutcliffe, B. T. Molecular structure and the born—Oppenheimer approximation. *Chem. Phys. Lett.* **45,** 393–398 (1977).

83. Gane, P. J. & Chan, A. W. E. Molecular fields in ligand discovery. *Methods Mol. Biol.* **1008,** 479–499 (2013).

84. González, M. A. Force fields and molecular dynamics simulations. *Collect. SFN* **12,** 169–200 (2011).

85. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J. Phys. Chem. B* **102,** 3586–3616 (1998).

86. Leach, A. R. Molecular modelling : principles and applications. *Computers* **21,** 784 (2001).

87. Hopfinger, A. J. & Pearlstein, R. A. Molecular mechanics force-field parameterization procedures. *J. Comput. Chem.* **5,** 486–499 (1984).

88. Naeem, R. Lennard-Jones Potential. *Chemwiki* 3–6 (2013). doi:10.1098/rspa.1924.0082.

89. Johnson, J. K., Zollweg, J. A. & Gubbins, K. E. The lennard-jones equation of state revisited. *Mol. Phys.* **78,** 591–618 (1993).

90. Levitt, M. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170,** 723–764 (1983).

91. Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A. & Joannopoulos, J. D. Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* **64,** 1045–1097 (1992).

92. van der Spoel, D. *et al.* Gromacs User Manual version 4.0. *Optimization* 308 (2005). doi:10.1007/SpringerReference_28001

93. Wagoner, J. & Baker, N. A. Solvation forces on biomolecular structures: A comparison of explicit solvent and poisson-boltzmann models. *J. Comput. Chem.* **25,** 1623–1629 (2004).

94. Cramer, C. J. & Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **99,** 2161–2200 (1999).

95. Maroncelli, M. & Fleming, G. R. Computer simulation of the dynamics of aqueous solvation. *J. Chem. Phys.* **89,** 5044–5069 (1988).

96. Skyner, R. E., McDonagh, J. L., Groom, C. R., van Mourik, T. & Mitchell, J. B. O. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **17,** 6174–6191 (2015).

97. Makov, G. & Payne, M. Periodic boundary conditions in ab initio calculations. *Physical Review B* **51,** 4014–4022 (1995).

98. De Vries, S. J., Van Dijk, A. D. J. & Bonvin, A. M. J. J. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins Struct. Funct. Genet.* **63,** 479–489 (2006).

99. Fenwick, R. B., Esteban-Martín, S. & Salvatella, X. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur. Biophys. J.* **40,** 1339–1355 (2011).

100. Tozzini, V. Coarse-grained models for proteins. *Current Opinion in Structural Biology* **15,** 144–150 (2005).

101. Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *Journal of Chemical Physics* **139,** 090901 (2013).

102. Scott, K. A. *et al.* Coarse-Grained MD Simulations of Membrane Protein-Bilayer Self-Assembly. *Structure* **16,** 621–630 (2008).

103. Antonietti, M. & Förster, S. Vesicles and Liposomes: A Self-Assembly Principle Beyond Lipids. *Advanced Materials* **15,** 1323–1333 (2003).

104. Monticelli, L. *et al.* The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **4,** 819–834 (2008).

105. Marrink, S. J. & Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **42,** 6801–22 (2013).

106. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & De Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *J.*

Phys. Chem. B **111,** 7812–7824 (2007).

107. Hansson, T., Oostenbrink, C. & Van Gunsteren, W. Molecular dynamics simulations. *Current Opinion in Structural Biology* **12,** 190–196 (2002).

108. Christen, M. *et al.* The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry* **26,** 1719–1751 (2005).

109. Van Meer, G., Voelker, D. R. & Feigenson, G. W. Membrane lipids: Where they are and how they behave. *Nature Reviews Molecular Cell Biology* **9,** 112–124 (2008).

110. Neves, M. A. C., Totrov, M. & Abagyan, R. Docking and scoring with ICM: The benchmarking results and strategies for improvement. *J. Comput. Aided. Mol. Des.* **26,** 675–686 (2012).

111. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

112. Bressert, E. *SciPy and NumPy*. *Journal of Chemical Information and Modeling* **53,** (2013).

113. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9,** 99–104 (2007).

114. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

115. DeLano, W. L. The PyMOL Molecular Graphics System. *Schrödinger LLC wwwpymolorg* **Version 1.,** http://www.pymol.org (2002).

116. Balgavý, P. *et al.* Bilayer thickness and lipid interface area in unilamellar extruded 1,2-diacylphosphatidylcholine liposomes: A small-angle neutron scattering study. *Biochim. Biophys. Acta - Biomembr.* **1512,** 40–52 (2001).

117. Brumshtein, B., Wormald, M. R., Silman, I., Futerman, A. H. & Sussman, J. L. Structural comparison of differently glycosylated forms of acid-β-glucosidase, the defective enzyme in Gaucher disease. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62,** 1458–1465 (2006).

118. Sujatha, M. S., Sasidhar, Y. U. & Balaji, P. V. Energetics of galactose- and glucose-aromatic amino acid interactions: Implications for binding in galactose-specific proteins. *Protein Sci.* **13,** 2502–2514 (2004).

119. Elgavish, S. & Shaanan, B. Lectin-carbohydrate interactions: Different folds,

common recognition principles. *Trends in Biochemical Sciences* **22,** 462–467 (1997).

120. Sundari, C. S. & Balasubramanian, D. Hydrophobic surfaces in saccharide chains. *Prog. Biophys. Mol. Biol.* **67,** 183–216 (1997).

121. Duan, X. & Quiocho, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **41,** 706–712 (2002).

122. Weis, W. I. & Drickamer, K. Structural Basis of Lectin-Carb Ohydrate Recognition. *Anna Rev. Biochen* **65,** 441–473 (1996).

123. Sa Miranda, M. C. *et al.* Heterogeneity in human acid β-glucosidase revealed by cellulose-acetate electrophoresis. *BBA - Gen. Subj.* **965,** 163–168 (1988).

124. Aerts, J. M. *et al.* Conditions affecting the activity of glucocerebrosidase purified from spleens of control subjects and patients with type 1 Gaucher disease. *Biochim. Biophys. Acta* **1041,** 55–63 (1990).

125. Qi, X. & Grabowski, G. a. Acid beta-glucosidase: intrinsic fluorescence and conformational changes induced by phospholipids and saposin C. *Biochemistry* **37,** 11544–11554 (1998).

126. Kobayashi, T. *et al.* Separation and characterization of late endosomal membrane domains. *J. Biol. Chem.* **277,** 32157–32164 (2002).

127. Alattia, J.-R., Shaw, J. E., Yip, C. M. & Prive, G. G. Molecular imaging of membrane interfaces reveals mode of beta-glucosidase activation by saposin C. *Proc. Natl. Acad. Sci.* **104,** 17394–17399 (2007).

128. Van Tilbeurgh, H., Bezzine, S., Cambillau, C., Verger, R. & Carrière, F. Colipase: Structure and interaction with pancreatic lipase. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* **1441,** 173–184 (1999).

129. Premkumar, L. *et al.* X-ray Structure of Human Acid-β-Glucosidase Covalently Bound to Conduritol-B-Epoxide. *J. Biol. Chem.* **280,** 23815–23819 (2005).

130. Qasba, P. K., Ramakrishnan, B. & Boeggeman, E. Substrate-induced conformational changes in glycosyltransferases. *Trends in Biochemical Sciences* **30,** 53–62 (2005).

131. Liou, B. *et al.* Analyses of variant acid β-glucosidases: Effects of Gaucher disease mutations. *J. Biol. Chem.* **281,** 4242–4253 (2006).

132. Salvioli, R., Tatti, M., Ciaffoni, F. & Vaccaro, A. M. Further studies on the reconstitution of glucosylceramidase activity by Sap C and anionic phospholipids. *FEBS Lett.* **472,** 17–21 (2000).

133. Ramanathan, A., Savol, A. J., Langmead, C. J., Agarwal, P. K. & Chennubhotla, C. S. Discovering conformational sub-states relevant to protein function. *PLoS One* **6**(1): e15827 (2011).

134. Ramanathan, A. & Savol, A. Protein Conformational Populations and. *Acc. Chem. Res.* **47,** 149–156 (2014).

135. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450,** 964–972 (2007).

136. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The Energy Landscapes and of Proteins Motions. *Science* **254,** 1598–1603 (1991).

137. Henzler-Wildman, K. A. *et al.* A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450,** 913–916 (2007).

138. Costa, M. G. S., Batista, P. R., Bisch, P. M. & Perahia, D. Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theory Comput.* **11,** 2755–2767 (2015).

139. Ramanathan, A. & Agarwal, P. K. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. *PLoS Biol.* **9,** (2011).

140. Prinz, J. H. *et al.* Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134,** (2011).

141. Burger, V. M. *et al.* Quasi-anharmonic analysis reveals intermediate States in the nuclear co-activator receptor binding domain ensemble. *Pac. Symp. Biocomput.* **1,** 70–81 (2012).

142. Ramanathan, A. & Agarwal, P. K. Computational identification of slow conformational fluctuations in proteins. *J. Phys. Chem. B* **113,** 16669–16680 (2009).

143. Savol, A. J., Burger, V. M., Agarwal, P. K., Ramanathan, A. & Chennubhotla, C. S. QAARM: Quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. *Bioinformatics* **27,** (2011).

144. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins Struct. Funct. Bioinforma.* **17,** 412–425 (1993).

145. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11,** 5525–5542 (2015).

146. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).

147. Noe, F., Schutte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci.* **106,** 19011–19016 (2009).

148. Northrup, S. H., Pear, M. R., McCammon, J. A., Karplus, M. & Takano, T. Internal mobility of ferrocytochrome c. *Nature* **287,** 659–660 (1980).

149. Ichiye, T. & Karplus, M. Anisotropy and Anharmonicity of Atomic Fluctuations in Proteins: Implications for X-ray Analysis. *Biochemistry* **27,** 3487–3497 (1988).

150. Ramanathan, A., Savol, A. J., Agarwal, P. K. & Chennubhotla, C. S. Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: Application to enzyme adenylate kinase. *Proteins Struct. Funct. Bioinforma.* **80,** 2536–2551 (2012).

151. Hyvärinen, A., Karhunen, J. & Oja, E. Independent component analysis. *John Wiley Sons* 481 (2001). doi:10.1255/nirn.1073

152. Molgedey, L. & Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **72,** 3634–3637 (1994).

153. Georgiev, P. & Chichocki, A. Robust independent component analysis via time-delayed cumulant functions. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **E86–A,** 573–579 (2003).

154. Forgy, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21,** 768–769 (1965).

155. Deuflhard, P. & Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **398,** 161–184 (2005).

156. Weber, M. & Kube, S. Robust Perron Cluster Analysis for Various Applications in Computational Life Science. *Lect. Notes Comput. Sci.* **3695,** 57-66-- (2005).

157. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **7,** 1192–1219 (2009).

158. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction.

J. Chem. Phys. **139,** (2013).

159. Kluyver, T. *et al. Jupyter Notebooks—a publishing format for reproducible computational workflows*. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016). doi:10.3233/978-1-61499-649-1-87

160. Rossum, G. Van & Drake, F. L. The Python Library Reference. *October* 1–1144 (2010).

161. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32,** 2319–2327 (2011).

162. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2012).

163. Röblitz, S. & Weber, M. Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv. Data Anal. Classif.* **7,** 147–179 (2013).